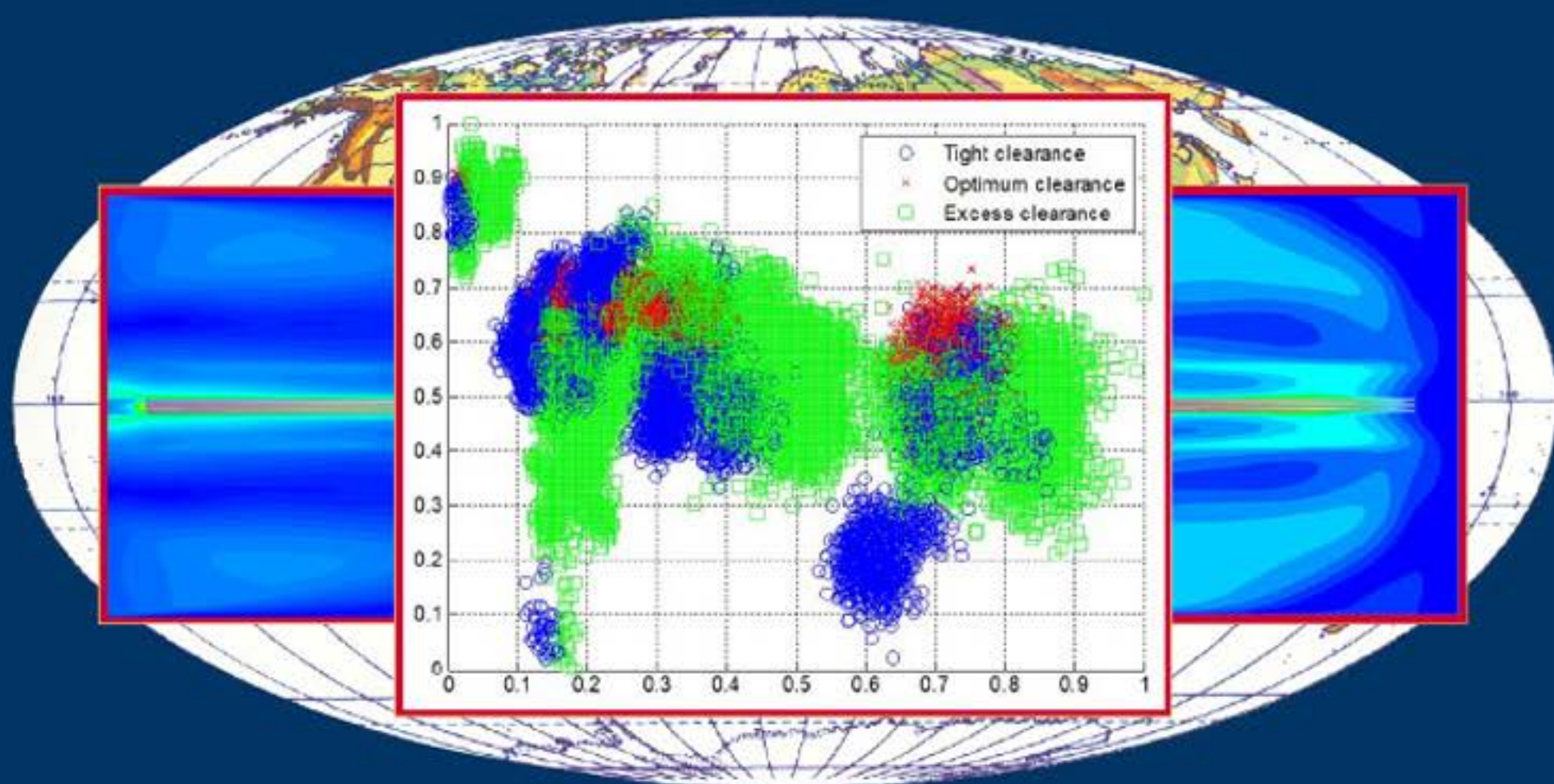


Vol. 23. No 4, 2021

ISSN 1507-2711
Cena: 25 zł

EKSPLOATACJA I NIEZAWODNOŚĆ

MAINTENANCE AND RELIABILITY



Polskie Naukowo Techniczne Towarzystwo Eksploatacyjne
Warszawa

Polish Maintenance Society
Warsaw

TABLE OF CONTENTS

| | |
|--|------------|
| Abolghasem Nobakhti, Sadigh Raissi, Kaveh Khalili Damghani, Roya Soltani | |
| Dynamic reliability assessment of a complex recovery system using fault tree, fuzzy inference and discrete event simulation | 593 |
| Łukasz Rymaniak, Jerzy Merksiz, Natalia Szymlet, Michalina Kamińska, Sylwester Weymann | |
| Use of emission indicators related to CO₂ emissions in the ecological assessment of an agricultural tractor | 605 |
| Guoxiao Zheng, Weifang Sun, Hao Zhang, Yuqing Zhou, Chen Gao | |
| Tool wear condition monitoring in milling process based on data fusion enhanced long short-term memory network under different cutting conditions | 612 |
| Przemysław Kowalak, Jarosław Myśków, Tomasz Tuński, Dariusz Bykowski, Tadeusz Borkowski | |
| A method for assessing of ship fuel system failures resulting from fuel changeover imposed by environmental requirements..... | 619 |
| Hao Lyu, Shuai Wang, Xiaowen Zhang, Zaiyou Yang, Michael Pecht | |
| Reliability modeling for dependent competing failure processes with phase-type distribution considering changing degradation rate..... | 628 |
| Jarosław Mamala, Mariusz Graba, Andrzej Bieniek, Krzysztof Prażnowski, Andrzej Augustynowicz, Michał Śmieja | |
| Study of energy consumption of a hybrid vehicle in real-world conditions | 636 |
| Haiyang Che, Shengkui Zeng, Qidong You, Yueheng Song, and Jianbin Guo | |
| A fault tree-based approach for aviation risk analysis considering mental workload overload | 646 |
| Paweł Gołda, Tomasz Zawisza, Mariusz Izdebski | |
| Evaluation of efficiency and reliability of airport processes using simulation tools | 659 |
| Thi-Phuong Nguyen, Yi-Kuei Lin | |
| Investigation of the influence of transit time on a multistate transportation network in tourism | 670 |
| Jakub Lewandowski, Stanisław Młynarski, Robert Pilch, Maksymilian Smolnik, Jan Szybka, Grzegorz Wiazania | |
| An evaluation method of preventive renewal strategies of railway vehicles selected parts | 678 |
| Chenchen Wu, Hongchun Sun, Senmiao Lin, Sheng Gao | |
| Remaining useful life prediction of bearings with different failure types based on multi-feature and deep convolution transfer learning..... | 685 |
| Katarzyna Antosz, Małgorzata Jasiulewicz-Kaczmarek, Łukasz Paśko, Chao Zhang, Shaoping Wang | |
| Application of machine learning and rough set theory in lean maintenance decision support system development | 695 |
| Konrad Lewczuk | |
| The study on the automated storage and retrieval system dependability | 709 |
| Gediminas Vaičiūnas, Stasys Steišūnas, Gintautas Bureika | |
| Specification of estimation of a passenger car ride smoothness under various exploitation conditions | 719 |
| Lijun Shang, Haibin Wang, Cang Wu, Zhiqiang Cai | |
| The post-warranty random maintenance policies for the product with random working cycles | 726 |
| Łukasz Jedliński | |
| Influence of the movement of involute profile gears along the off-line of action on the gear tooth position along the line of action direction | 736 |
| Yi Lyu, Yijie Jiang, Qichen Zhang, Ci Chen | |
| Remaining useful life prediction with insufficient degradation data based on deep learning approach | 745 |
| Karol Andrzejczak, Lech Bukowski | |
| A method for estimating the probability distribution of the lifetime for new technical equipment based on expert judgement..... | 757 |
| Edward Michlowicz, Jerzy Wojciechowski | |
| A method for evaluating and upgrading systems with parallel structures with forced redundancy | 770 |
| Edward Kozłowski, Katarzyna Antosz, Dariusz Mazurkiewicz, Jarosław Sęp, Tomasz Żabiński | |
| Integrating advanced measurement and signal processing for reliability decision-making | 777 |

Dynamic reliability assessment of a complex recovery system using fault tree, fuzzy inference and discrete event simulation

Abolghasem Nobakhti^a, Sadigh Raissi^{a,b*}, Kaveh Khalili Damghani^a, Roya Soltani^c

^aSchool of Industrial Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran

^bResearch Center for Modeling and Optimization in Science and Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran

^cDepartment of Industrial Engineering, KHATAM University, Tehran, Iran

Indexed by:




Highlights

- Focus on reliability assessment under dynamic operations and lack of historical data.
- FTA helped to estimate reliability fitness function for each alternative.
- The Mamdani fuzzy inference handled multi-attribute failure risks using a questionnaire.
- The FOVs acts better at dynamic conditions in terms of discrete-event simulation.

Abstract

Any failure on the recovery system will cause a lot of environmental damage as well as energy loss. Hereof two types of alternatives; fast opening valve system (FOVS) and seal drum system (SDS) may be installed. The focus of this article is on the decision stage to choose the most preferred option in terms of reliability assessment. The major challenge in the research problem is on changing the pressure and temperature during operational cycles, which significantly affect the reliability. In addition, the lack of historical data complicates the reliability assessment method. Hence, we proposed a hybrid approach using fault tree analysis (FTA) and the Mamdani fuzzy inference to estimate reliability response as a function of a few frequently operating pressure and temperature. Also, discrete-event simulation helped us to evaluate the system reliability at different operating conditions. The comparisons reveals that the FOVs outperforms on average of 22.4% than the SDS and it is recommended for putting into practice for purchasing.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

dynamic reliability assessment, Fault Tree Analysis (FTA), Mamdani Fuzzy Inference Method; discrete event simulation, flare gas recovery system.

1. Introduction

Today, saving energy and preventing environmental pollution caused by burnt fossil fuels are two important issues in the refinery equipment selection process. It is no longer time for flare gases to be burned in refineries and, in addition to wasting a good energy source, to injury the environment.

In almost all societies, tackling the significant environmental damage caused by fossil fuel emissions is on the agenda of senior executives. They must use all means in strategic decision-making to reduce environmental losses to save the future. Countries participating in the Paris climate agreement, developed under the United Nations Framework Convention on Climate Change (UNFCCC), are bound to take actions to reduce their greenhouse gas emissions to meet a nationally determined contribution (NDC). According to [8], flaring reduction plays a major part in facilitating the reduction of emissions and reaching the targeted NDC. To this end, refinery industries are setting aside portions of their budgets to expand refineries lacking flare gas recovery units (FGRUs) to recover a large portion of flare gas, making it available for energy production etc.

It is vital that FGRU has a continuous operation as it prevents extra emission to the atmosphere and returns large benefits, so a proper

safety sub-system is required to prevent failures caused by the out-of-range pressure and temperature of the gas. Therefore, decision-makers are facing the daunting task of choosing among the proposed plans for the FGRU safety structures. The chosen alternative must be fully justified in terms of resilience against the volatile operating conditions, thus the need to make a comprehensive prediction of system reliability.

The selected alternative will be operating for more than two decades and making the wrong choice can lead to huge financial losses or a large amount of pollution because of more failures. Since the decision-making is performed in the pre-installation and purchasing stage, failure data are not available. Besides that, precision in reliability prediction needs the consideration of alterations in reliability value caused by operational conditions and contributing factors. The generated reliability values must be responsive towards more than only time which is the output of the traditional reliability methods. Demonstration of the changes of reliability versus certain contributing factors requires proper initial data able to describe such changes and a proper technique to process such data.

Expert elicitation is widely used in papers to compensate for the void of unavailable information when prediction is needed and has

(*) Corresponding author.

E-mail addresses: A. Nobakhti - st_a_nobakhti@azad.ac.ir, S. Raissi - Raissi@azad.ac.ir, K. K. Damghani - k_khalili@azad.ac.ir, R. Soltani - r.soltani@khatam.ac.ir

the flexibility to provide a researcher with the desired type of data. The failure probability of components can be assessed in different circumstances using the opinion of a group of experts. Changes in failure probability versus contributing factors can help calculate the relevant changes of safety system reliability that is the main requirement for a prognostic study. Quantification techniques generate probabilities from linguistic possibilities so that calculations are performed. Data are gathered for the certain key components whose failure will cause the failure of the system and to identify those components a functional and physical breakdown is required.

Fault tree analysis (FTA) can demonstrate the functional and physical breakdowns of the system and can be inserted with different types of data. This provides the possibility of branching and dividing failure causes and detecting the groups of components whose failures cause a major system failure. FTA uses logical gates such as “AND” and “OR” to describe the effects of components’ failure on system breakdown using the Boolean algebra and has widely been used for safety, risk, and reliability analysis. However, as for most techniques, FTA has its own limitations. For instance, FTA’s routine calculation methods can’t describe the changes of the output versus changes of the basic event values. Time is one of the factors able to change the basic event values and in certain cases, the similarities of the output values in a specified period of time would lead to choosing the wrong alternative. Yet, time isn’t the only factor that influences FTA’s output. For systems operating under volatile conditions, stress factors also affect FTA’s output but their influence can’t be modeled.

FTA receives numerical data, so in order to quantify the linguistic data gathered from expert elicitation, the fuzzy inference system (FIS) is a powerful technique to be used for quantification. This helps describe changes in basic event probabilities with the changes of stress factor levels. So, a fusion between FIS and FTA leads to having a responsive FTA output. With this being said, neither a lot of reported research exists on the development of a technique for FTA enrichment nor has there been any work on the combination of FIS and FTA.

The main goal of this paper is to develop a practical technique to compute the changing output of the FTA, which is reliability in this paper, versus contributing factors so that an accurate prediction of performance is made. Performance predictions help in making a justified choice between the proposed alternatives in the absence of historical data. The proposed fuzzy interface combinatorial usage of FTA and FIS (FIFTA) is an innovative approach proposed here to tackle reliability estimation in the absence of historical data. Also, for more accurate calculation, a combination of FTA and discrete-event simulation is used so that with the help of this alternative method, the reliability of the two systems can be calculated.

These two methods can also graphically illustrate reliability response surface as a function of the relevant covariates to provide insights especially for decision makings in the purchasing stage. It provides an applicable method for a facile computational prediction of future performances that aims to replace the usage of failure rates by a combination of instructed expert elicitation and fuzzy inference system and discrete event simulation.

In section 2, there is a review of papers with similar cases. Section 3 provides information about the proposed alternatives for the safety sub-systems. In section 4, the challenges of calculating valid reliability for the mentioned alternatives are discussed. The proposed method for this is found in sections 5 and 6. Two methods are implemented on the alternatives, and the results are presented in section 7 where further discussion is also made for more clarification. A conclusion is made in section 8.

2. Literature review

Reliability engineering is a major sub-discipline for systems engineers to assess the probability of surviving a system over time. This method focuses on lifetime evaluation under stated conditions for a specified period of time. Various researchers have tried to provide ef-

ficient methods for estimating system reliability based on empirical data. Some of them proposed aggregate method of selecting a theoretical distribution for empirical data [19]. They applied three criteria for assessing the quality of the goodness of fit.

If the operating conditions change, then the reliability analysis will be a difficult task, which is a matter of dynamic reliability. In dynamic reliability analysis, a set of the mathematical framework is presented which has the capability of handling interactions among components and process variables. In principle, they constitute a more realistic modeling of systems for the purposes of reliability, risk, and safety analysis. Dynamic reliability requires more sophisticated tools than non-dynamic reliability. Dynamic reliability needs to apply a more complicated mathematical methods approach takes into that account changes or evolution of the system structure.

Changes in process parameters may be random or deterministic. Indeed, reliability modeling of the former is far more difficult than the latter and is often accomplished by computer simulation techniques. Interested readers could refer to [11] for deterministic changes, [4] for stochastic changes and [18] for ranking defects.

Ambiguity and vagueness are issues that are caused by the unknown characteristics of the complex systems or insufficiency of historical failure data that leads to making rough estimations, hence increased error in the final results. Therefore, to minimize this error, fuzzy logic may be a proper alternative [16]. A combination of FTA and fuzzy logic would create the new (FFTA) technique that has widely been studied in recent years where expert elicitation is used to obtain the linguistic values as possibilities which are then transformed into quantitative probabilistic values for basic events of the fault tree. [15] Employed a combination of fuzzy logic and expert elicitation to deal with vagueness and subjectivity of the information and generated basic event failure probabilities without reliance on quantitative historical failure data and performed a sensitivity analysis using importance measuring. Yazdani et al. [21] used fault tree qualitative analysis technique to identify various potential causes of crude oil tank fire and explosion (COTFE) and used a hybrid approach of fuzzy set theory to quantify the COTFE fault tree; the results were compared with that of a conventional fault tree. Weak links were identified using importance measuring of basic events. [14] proposed a fuzzy-based reliability approach to deal with qualitative linguistic terms to evaluate the failure likelihood of basic events of nuclear power plant safety system; and validated the results by a benchmarking the generated failure probability to the actual failure probabilities collected from the operating experiences of the David-Besse design of the Babcock and Wilcox reactor protection system.

Certain papers went further and tried to improve the elicitations and didn’t stop on a sole reliance on raw opinions. Baig et al. [3] used corrosion simulation software and provided the experts with the obtained results to improve the elicitations. They gathered information to estimate the failure probability of CO₂ transporting the pipeline using FTA. Attention to computer simulation in estimating reliability has been considered by various researchers in recent years. The reason for this is the existence of different random variables and the complexity of systems analysis by analytical methods. For example, we can refer to [1, 11], and [13], whose methods have been cited by many researchers.

In order to deal with the uncertainty in linguistic data, researchers have often recommended the use of fuzzy methods. A two-dimensional fuzzy fault tree analysis to incorporate hesitation factor for expert elicitation where linguistic terms were expressed with a degree of hesitation introduced by [20]. Through applying such a technique, the probability of chlorine release was estimated for Indian conditions.

In cases where historic data is insufficient but a failure rate may be obtained from a given static failure distribution that could satisfy the desirable accuracy, it is possible to obtain information from data banks like OREDA. Elsayed [6] performed a four-step procedure to estimate reliability with failure and repair data from OREDA and calculated availability and maintainability as well. Zhang et al. [22]

graded a floating offshore wind turbine (FOWT) system structurally and functionally, thereby assessing the sequentially dependent failures and redundancy failures using a dynamic fault tree. Reliability estimation was based on failure data achieved from OREDA. In order to nominate a diagnostic method and measuring the total predictive performance score, an integrated fuzzy DEMATEL-fuzzy analytic network approach presented in [12].

In the case study cited in the current research, none of the aforementioned alternatives for the recovery of flare gases were practically available, and empirical data on their performance were not available, so we had to use the experiences of technical experts in similar matters. This made the data collected via linguistic variables and we needed to use the appropriate tools for quantification to perform the calculations. Therefore, as a new initiative, a combination of the Mamdani Inference System; FIS and The Fault Tree Analysis; FTA methods has been used to investigate the various failure modes under different operating conditions. However, it has been used in several cases for approximation and estimation with different purposes. Azadeh et al. [2] used FIS as a means of approximation for human reasoning to provide knowledge for correct and timely diagnosis of pump failures. Choi et al. [5] used FIS in combination with relative risk score (RRS) as a new approach for liquid and gas pipeline risk assessment and proved that the new method provides more accurate results in comparison with the conventional method. Elvidge et al. [7] used Mamdani and Sugeno FIS as an alternative approach to qualitative risk matrix to handle multiple attribute risk problems with imprecise data. He found out that while Mamdani method is intuitive and well suited for human inputs, the Sugeno method is computationally more efficient and guarantees the continuity of the final risk output surface.

Nematkhah et al. [9] investigated some methodologies to how to decrease energy consumption and reduce the environmental pollution of flare systems. In this study, three different scenarios evaluated by the use of an environmental flow diagram in a gas refinery in southern Iran. The results showed that pressurizing gas and injecting it into oil wells is one of the best ways to reduce flames in the Feller gas system. [22] studied three different system configurations on flare gas recovery to evaluate the efficient system. In this study, systems with liquid ring compressors and aqueous amine solvents for the abatement of acid gases are used in a refinery complex. The results show that amine consumption in some configurations is much lower than in others.

Recently, two designs of flare gas recovery systems were designed and reliability was chosen as the deciding factor for comparing two systems. First, failure models of the two designs have been implemented. Second, a stochastic hybrid method is used to evaluate the probability of disaster in these failures [8].

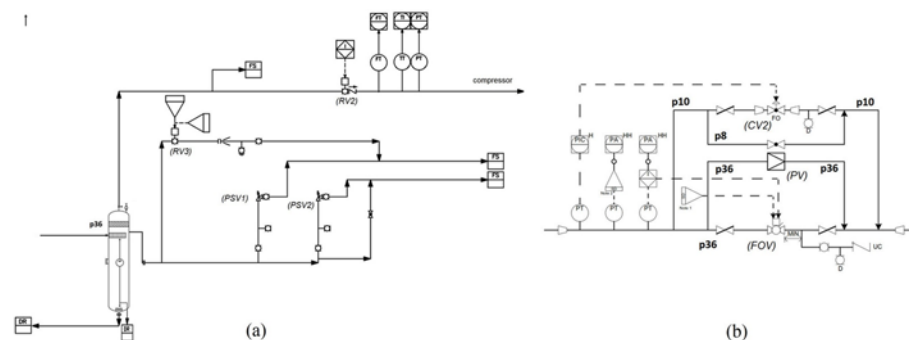


Fig. 1. A schematic view of the FOVS (a) and SDS safety subsystems (b)

3. System description

Two alternatives are proposed as FGRU safety sub-systems to keep it intact against out-of-range characteristics of passing gas. These alternatives have many similar but their main difference is in the pre-flaring section which can either be a fast-opening valve system (FOVS) and seal drum system (SDS). The relevant diagrams are depicted in Figure 1 as (a) and (b) respectively.

There are various incidents that can lead to damage or FGRU breakdown. A dangerous scenario may occur when out of control gas pressure or gas temperature happens. Three hazardous scenarios are discussed in section 5.3. The purpose of installing a safety system is to block the routs leading to FGRU to keep it intact and to open more capacity to the flaring tower to prevent piping ruptures.

There are pre-defined responses towards each scenario in each safety system that is initiated when dangerous temperature or pressure is detected by sensors and proper messages are sent to the valve actuators. The actuators receive the signals from sensors and open or close a valve's body; thereby directing the gas with dangerous temperature or pressure level to the flaring tower. If the safety subsystem, fails to respond towards a dangerous scenario, not only risky occur to FGRU but the safety subsystem itself is likely to get damaged.

A general view of the FGRU depicted in Figure 2. Hence gas enters from the flare header to the safety system and is directed in a proper volume to the compressor to get prepared for recovery. The route leading to the recovery section is called the 'vertical route'. The extra gas or gas with dangerous characteristics will be transferred through the 'horizontal route' to be burnt in the flaring tower. The components' names and symbols are provided in Table 1.

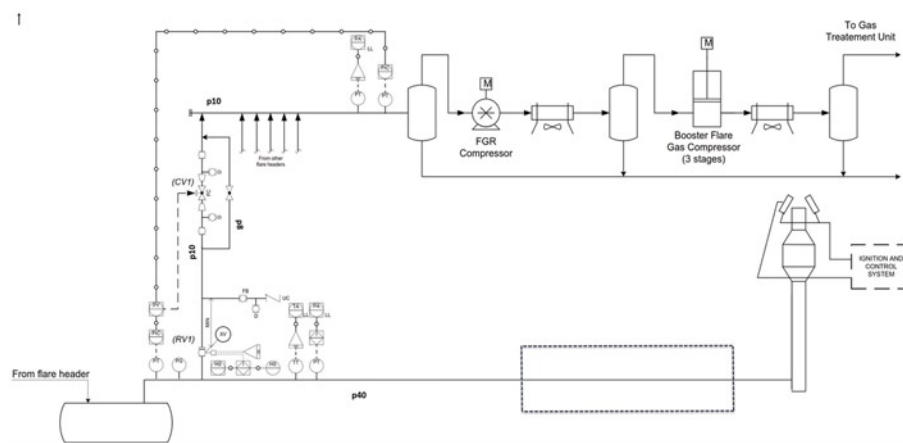


Fig. 2. Schematic view of flare gas recovery unit (FGRU)

There are 3 main components in a vertical route that prevent the entrance of gas with dangerous characteristics to the compressor which are a rotary valve (RV1), a control valve (CV1), RV1's task is to close with sensors' message and CV1 must close when a difference of pressure is detected between system entrance and the entrance to seal drum (SD).

The horizontal route leads to the flaring tower before which (in the area depicted with a dashed line) the safety system (FOVS or SDS) must be installed here to react to the signals sent from sensors. This part of the safety system opens more capacity to the pipes, so the extra gas is emitted without causing any damage or helps direct some extra gas to the flaring tower to prevent flashbacks. Flashback is the result of very low pressure in the horizontal route that will reverse the direction of the gas and damage pipes and components.

Of the two safety systems, SDS is a collection of SD and the accompanying valves which are two rotary valves (RV2 & 3) and two pressure safety valves (PSV 1& 2). SD in SDS, contains a proper level of water to keeps gas flow in a single direction (from inlet to the outlet) which is helpful in preventing flashback making it quite useful for implementation in the pre-flaring section. SD also prevents gas outlet until the pressure reaches a desired, and often predetermined pressure.

On the other hand, FOVS remains a collection of valves that respond to different scenarios by a harmonious function of sensors to make a safe passage for gas in a fashion that damages are prevented to the piping systems or to the FGRU. It comprises of a control valve (CV2) and a reserve line for when CV2 is being repaired, a pin valve (PV) and a fast-opening valve (FOV). When pressure increases in the horizontal route, valves in this structure will unlock one by one to provide more capacity for gas to be released into the flaring tower.

4. Problem statement

The valuable components and repair costs of FGRU raises the imperative of the fully justified selection of a safety subsystem, resilient against the volatile operating conditions, to protect the FGRU against gas with dangerous characteristics. The resulting reduced damages to FGRU, apart from expenses, helps to minimize the emitted gas to the atmosphere, facilitating meeting NDC.

Of the two suggested alternatives for the safety subsystem are FOVS and SDS. The one with higher reliability and consequently fewer failures should be chosen to decrease FGRU damages. FOVS or SDS will be the pre-flaring section of the safety subsystem whose components interact with other components of the other sections so, making an isolated reliability assessment of them without considering their interrelations wouldn't be valid. So, to compare them in term of reliability, the performance of the whole safety system must be assessed when either of them installed.

The traditional reliability methods only considered the dependency on time which overlooked the environmental factors. Using such results leads to having to tackle unpredicted failures in such a volatile environment and the objective is to obtain the reliability of the subsystem when it is exposed to different operating conditions and different scenarios.

Generating reliability values versus the three contributing factors of the studied case (time, pressure, and temperature), requires a specific type of data able to associate an operating condition to a failure probability value. In other words, a function is required with a domain that consists of a space made up of three axes of time, pressure and the temperature limited to their boundaries (i.e. maximum, and minimum levels of contributing factors). The codomain is a value between (0, 1) that describes a failure probability. In other words, a type of failure data is to be provided for each component that describes its endurance under a certain operating condition. Obtaining such data isn't possible through measurement because the alternatives haven't been installed yet, and there is no such data in the data-banks.

When experiencing the need to making calculations for a system in its pre-installation stage, the available type of data are failure rates gathered with the assumption of a stable failure distribution from other similar systems. Reliability calculations based on failure rates show only reliability changes versus time and the assumption of a stable failure distribution neglects the effects of the stress factors. It is

professionally recognized that the failure distributions' scale changes with the presence of a stress factor whose level is higher than that the operating condition. This alters the area under the distribution function and consequently changes the reliability values.

Apart from the need to gather a type of data that can describe the simultaneous presence of the contributing factors, a technique is required to process the data so that it is available to be used in the fault tree. It is intended to generate a response surface for reliability to study its changes versus contributing factors. Data is gathered using a designed questionnaire and the utilized technique is FIS, both of which are explained in the next section.

5. The proposed method to estimate system reliability surface

In order to estimate the recovery unit reliability as a function of operating condition, dynamic fault tree analysis (FTA) fixed as the main core of the estimation. Due to lack of historical data, expert judgment is used on the failure likelihood of each component at different operating circumstances. Then Mamdani fuzzy inference method is applied to quantify the linguistic data and to generate different points to draw the response surface for each alternative in a four-dimension space. A general overview of the proposed method is depicted in Figure 3.

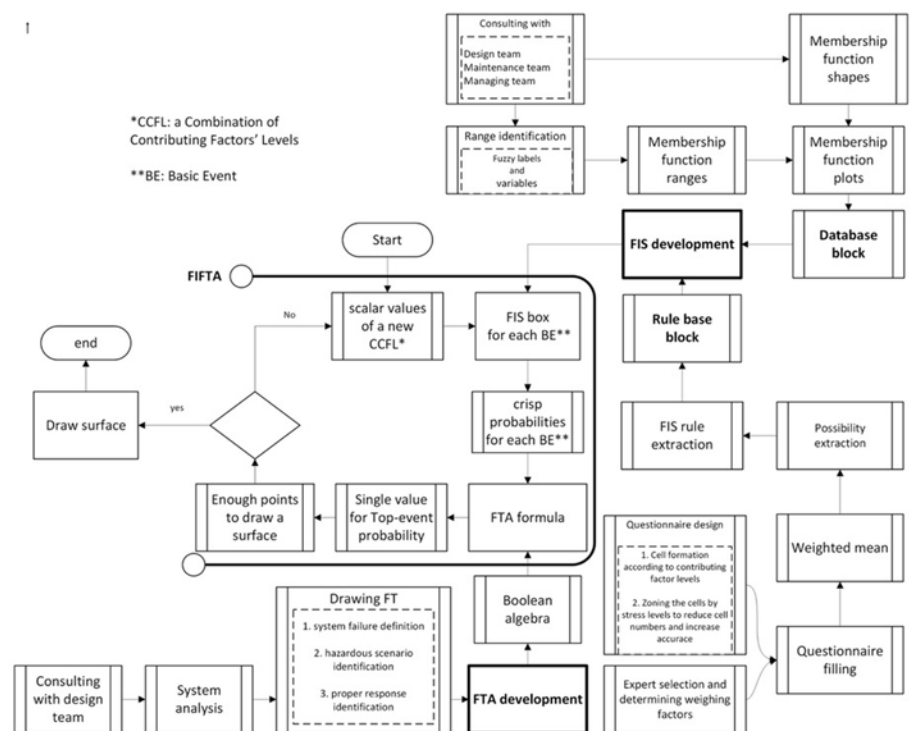


Fig. 3. The proposed fuzzy inference fault tree analysis (FIFTA) methodology

Due to a lack of historical data in purchasing stage, we prepared a verified reliable questionnaire (Table 2) to analyze each alternative component's breakdown likelihood over different process conditions based on the experts' opinions. Here temperature and pressure are deduced as the main contributive factors on the components' failure. To overcome the ambiguity, arouse from linguistic terms we converted all despondence via normalized fuzzy sets.

The gathered data presented component failure possibilities in association with temperature and pressure levels. Any data point reveals an expected prior possibility of a component lifetime at a given temperature and pressure using a triplet of (time, pressure, temperature). The purpose is to quantification that possibility so that a component failure probability response surface is drawn. The surface will associate each component breakdown probability with an operating condition. Converting possibility into probability requires quantification

performed through FIS. But first fault tree (FT) should be drawn to model the failures. 5.1 to 5.3 present the section of Figure 3 that is related to drawing FT. The section of Figure 3 that concerns FIS development is presented in 5.4 to 5.6 and finally, FIFTA is presented in 5.7.

5.1. System failures

In order to be able to draw FT, an explicit definition of failure is required. For that, the structural and functional breakdown of the system should be examined. The structural breakdown of the system indicates that the critical components of the FGUR are: pipes, sensors, valves, and compressors. The functional breakdown of the system indicates that gas is directed by pipes into compressor or flaring section, sensors detect temporal characteristics of gas and send signals to valves when the gas with out-of-range pressure or temperature enters, valves change the route of gas and open more exit capacity so that compressors or pipes are not damaged. Compressors alter the characteristics of the gas so that it is ready to be recovered.

System failure occurs if the gas route isn't altered because of valve failures or gas isn't directed toward to the compressor because of piping damages. Valves fail under the effect of changing pressure and temperature that accelerate valve body degradation. If valves fail, pipes and compressors are exposed to the danger of getting damaged by a hazardous scenario (5.3). Therefore, a failure definition can be presented as follows:

1. Valves degrade gradually to the point of not being able to function in demand.
2. A hazardous scenario occurs i.e. gas with an out of the standard level of temperature or pressure enters.
3. Automated system fails to respond i.e. gas with out-of-range pressure or temperature isn't directed appropriately because of valve failures.
4. Gas causes damages to the compressor or critical pipes, and the system fails.

This definition helps us divide basic events of FT and form the branches. These four segments occur respectively but FTA logical gates can't enforce the order of occurrence. DFTA gates can't be used because failure rates are required for solution and they are unapt for this study as there is the need to assess multi-dimensional data; so, inhibit gate is inevitably used to describe the relation between them. The second segment of failure definition is not a failure but an event, but it is presented in the model and its probability is considered the percentage of time that it happens (each percentage is presented in 5.3).

Valves' failure is caused by the changing pressure and temperature so failure data is gathered using the questionnaire in Table 2. These data will be the basis of FIFTA study where we insert different numerical levels for pressures, temperatures and times into FIFTA to study the changes of failure probabilities of valves and the whole system. But pipe and compressor failure probabilities are obtained using a different questionnaire where experts are only asked to specify the failure possibility of the components under one of the three hazardous scenarios. This is due to the fact that the cause of their failure is the occurrence of a hazardous scenario when there is no proper response. Defuzzification of these possibilities is performed using the method described in [10] for each scenario. The obtained probabilities are considered as a constant in FTA formula and the basic events describing their damages are not a part of FIFTA process.

It should be stated that the independent failures of pipes and compressor (i.e. failures caused by initial defects, by degradation, by faulty design, etc.) are not considered here and it is assumed that they

will remain intact in normal conditions during the predicted lifetime because of the sufficient protective measures and high-quality materials. Also, sensor failures aren't taken into account since changes in temperature or pressure have such a small effect on them that it can be neglected and since they are of high-quality materials, their independent failures are omitted from calculations.

5.2. Constructing dynamic fault tree analysis

Fault Tree; FT is constructed for both systems according to the above-mentioned failure definition. The first levels of this diagram are presented in Figure 4-a for FOVS and in Figure 4-b for SDS.

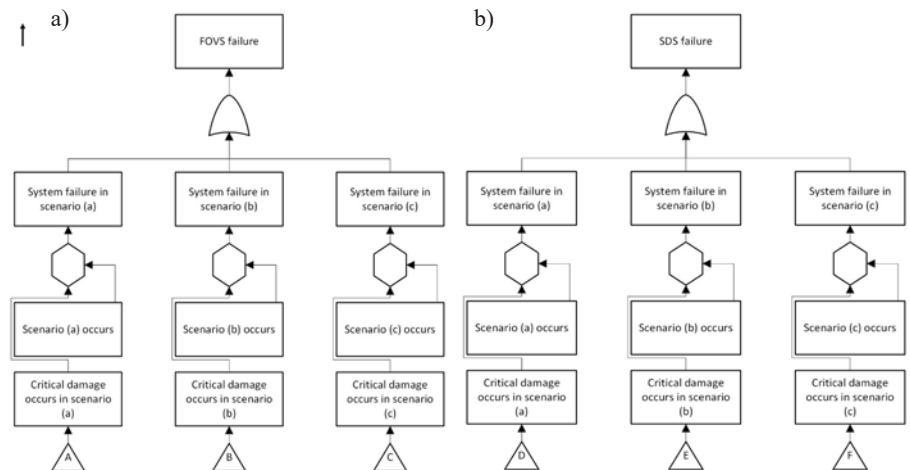


Fig. 4. The main failure causes for FOVS (a) and SDS (b)

5.3. Most common hazardous scenarios

In order to examine fault tree in dynamic circumstances, three more probable extreme operational conditions examined in this research, they called hereinafter as:

Scenario a: Examining the failure of the system at high pressure operating conditions with a chance of 33% according to the historical data.

Scenario b: Examining at Low pressure) with a probability of 29% in occurrence.

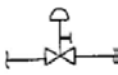
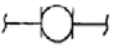




Scenario c: Examining at low temperature (22% occurrence).

Since there are two alternatives of FOVS and SDS for comparison at the above-mentioned three scenarios, six fault tree diagrams should be constructed. Figure 5 illustrates one of them as a sample. Interested readers can receive other diagrams by their request to authors.

5.4. Questionnaire cell formation

As mentioned earlier, data should be gathered with a properly designed questionnaire. In the designed questionnaire, experts are asked to express their opinion about the failure possibility of a component that ensures a certain operating condition created by contributing factors. For example, condition 1 is when a component is in its early age period, and endures a low pressure, and a low temperature, first cell of the questionnaire, and the expert provides a linguistic value in that cell using a fuzzy label like 'low' to describe failure possibility of the component in that condition. This linguistic data contains 3 input variables (i.e. time, pressure and temperature) and 1 output variable (failure possibility) giving it multiple dimensions. The purpose of gathering data in this manner is to study failures in each operating condition so that the whole system can be studied under each condition. As a result, the questionnaire should be designed in a manner that every cell represents an operating condition. In each cell, the expert describes the failure likelihood of the component in that condition. Table 2 shows the design questionnaire.

Table 1. Valves and their types of failures

| Name | symbol | Abb. | Used in | Failure type |
|-----------------------|---|------|---------|-----------------|
| Control valve |  | CV1 | Fig. 2 | Fail to close |
| | | CV2 | Fig. 4 | Fail to open |
| Rotary valve |  | RV1 | Fig. 2 | Fail to close |
| | | RV2 | Fig. 3 | Fail to close |
| | | RV3 | Fig. 3 | Fail to open |
| Pin valve |  | PV | Fig. 4 | Fail to rupture |
| Fast opening valve |  | FOV | Fig. 4 | Fail to open |
| Pressure safety valve |  | PSV1 | Fig. 3 | Fail to open |
| | | PSV2 | Fig. 3 | Fail to open |
| Spare Globe valve |  | SP1 | | Fail to close |
| | | SP2 | | Fail to open |
| | | SP3 | | Fail to open |
| | | SP4 | | Fail to open |

5.5. The Mamdani fuzzy inference system

Failure possibility examined based on information gathered from qualified experts. Their judgments requested different operation conditions using lingual terms, which modeled by fuzzy numbers. Some researchers have used the method of fuzzy inference in the oil and gas and petrochemical industries for risk analysis [6]. Hence Mamdani FIS is applied to create a control system by synthesizing a set of linguistic control rules obtained from experienced human operators. In a Mamdani system, the output of each rule is described by a fuzzy set. Since Mamdani systems have more intuitive and easier-to-understand rule bases, they are well-suited to expert system applications where the rules are created from human expert knowledge, such as medical diagnostics. This technique generates a numerical value i.e., failure probability. E.g. it is required to know the failure probability of RV that has operated for 2 years when it endures a pressure of 50 bars and a temperature of 0°C. Each cell in the questionnaire describes this operating condition to a degree between (0, 1) i.e. membership function. The opinions for each operating condition are aggregated based on the membership functions of each cell to generate a failure probability. The generated probability by FIS suggests for the above example, that there is a 0.02 chance of failure for RV in that condition. Since there are 9 types of valves (Table 1), and opinions vary about their failure likelihood, a FIS should be developed for each of them.

Table 2. Weighted mean of failure possibilities judged by 12 experts

| | Time | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--------|----------|---|---|------|---|---|------|---|---|----------|----|----|------|----|---|------|----|----|----------|----|----|------|----|----|------|----|----|
| | ↓ | | | | | | | | | – | | | | | | | | | ↑ | | | | | | | | |
| | Pressure | | | | | | | | | Pressure | | | | | | | | | Pressure | | | | | | | | |
| | ↓ | | | – | | | ↑ | | | ↓ | | | – | | | ↑ | | | ↓ | | | – | | | ↑ | | |
| | temp | | | temp | | | temp | | | temp | | | temp | | | temp | | | temp | | | temp | | | temp | | |
| | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ | ↓ | – | ↑ |
| stress | B | G | G | B | G | G | O | Y | Y | Y | B | B | Y | B | B | R | O | O | O | Y | Y | O | Y | Y | R | R | R |
| RV1 | 10 | 3 | 3 | 10 | 3 | 3 | 17 | 9 | 9 | 13 | 11 | 11 | 11 | 11 | 9 | 22 | 17 | 17 | 18 | 11 | 11 | 18 | 13 | 11 | 24 | 24 | 23 |

↑: high --: medium ↓: low

FIS Rules for the basic event, representing RV1 fail to act:

1. If (Time is low) and (Temperature is low) and (Pressure is low) then (possibility is 10) (1)
2. If (Time is low) and (Temperature is medium) and (Pressure is low) then (possibility is 3) (2)
3. If (Time is low) and (Temperature is high) and (Pressure is low) then (possibility is 3) (3)
4. If (Time is low) and (Temperature is low) and (Pressure is medium) then (possibility is 10) (4)
5. If (Time is low) and (Temperature is medium) and (Pressure is medium) then (possibility is 3) (5)
6. If (Time is low) and (Temperature is high) and (Pressure is medium) then (possibility is 3) (6)
7. If (Time is low) and (Temperature is low) and (Pressure is high) then (possibility is 17) (7)
8. If (Time is low) and (Temperature is medium) and (Pressure is high) then (possibility is 9) (8)
9. If (Time is low) and (Temperature is high) and (Pressure is high) then (possibility is 9) (9)
10. If (Time is medium) and (Temperature is low) and (Pressure is low) then (possibility is 13) (10)
11. If (Time is medium) and (Temperature is medium) and (Pressure is low) then (possibility is 11) (11)
12. If (Time is medium) and (Temperature is high) and (Pressure is low) then (possibility is 11) (12)
13. If (Time is medium) and (Temperature is low) and (Pressure is medium) then (possibility is 11) (13)
14. If (Time is medium) and (Temperature is medium) and (Pressure is medium) then (possibility is 11) (14)
15. If (Time is medium) and (Temperature is high) and (Pressure is medium) then (possibility is 9) (15)
16. If (Time is medium) and (Temperature is low) and (Pressure is high) then (possibility is 22) (16)
17. If (Time is medium) and (Temperature is medium) and (Pressure is high) then (possibility is 17) (17)
18. If (Time is medium) and (Temperature is high) and (Pressure is high) then (possibility is 17) (18)
19. If (Time is high) and (Temperature is low) and (Pressure is low) then (possibility is 18) (19)
20. If (Time is high) and (Temperature is medium) and (Pressure is low) then (possibility is 11) (20)
21. If (Time is high) and (Temperature is high) and (Pressure is low) then (possibility is 11) (21)
22. If (Time is high) and (Temperature is low) and (Pressure is medium) then (possibility is 18) (22)
23. If (Time is high) and (Temperature is medium) and (Pressure is medium) then (possibility is 13) (23)
24. If (Time is high) and (Temperature is high) and (Pressure is medium) then (possibility is 11) (24)
25. If (Time is high) and (Temperature is low) and (Pressure is high) then (possibility is 24) (25)
26. If (Time is high) and (Temperature is medium) and (Pressure is high) then (possibility is 24) (26)
27. If (Time is high) and (Temperature is high) and (Pressure is high) then (possibility is 23) (27)

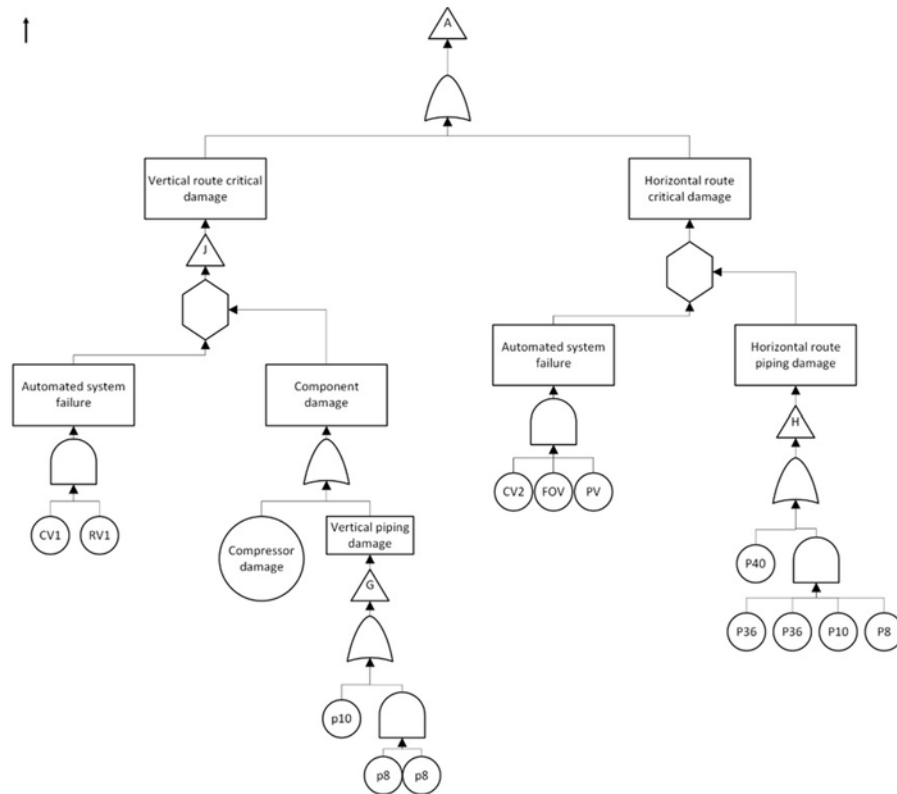


Fig. 5. Failure tree diagram for FOVS at the high-pressure scenario

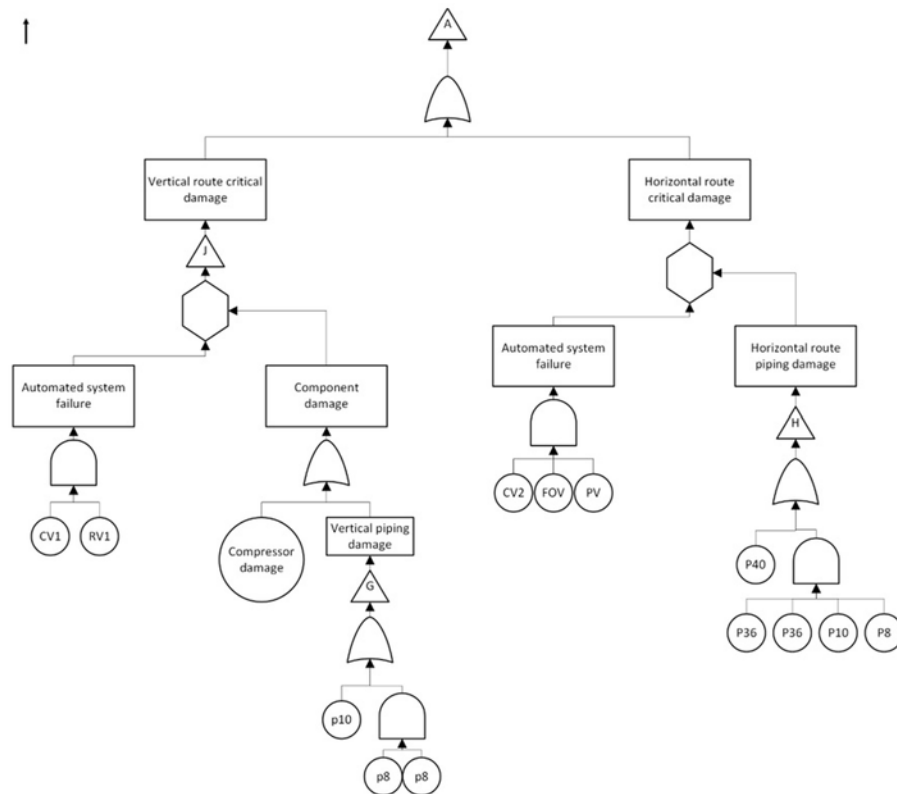


Fig. 6. The proposed Mamdani fuzzy inference system

The FIS has 5 functional blocks to measure data with multiple input and output variables. Of these 5 blocks, database block and rule base block store predetermined data and if these blocks are formed, the others will perform the quantification. The formation of these two is performed as follows.

When the input and output variables are identified, their range of variation is specified. Then the desired number of fuzzy labels (3, 5, 10, etc.) divide the variable's variation range. Here, 3 labels namely high, medium and low are used for each input variable (whose combination builds up questionnaire cells) and 25 labels for the output variable (labels of the opinions). Fuzzy membership functions are defined for each fuzzy set which consists of the shape of each function (e.g. triangular) and its boundary. These data are stored in a database block whose formation is depicted at the top-right hand side of Figure 3. Figure 6 contains the shapes and boundaries of the membership functions for each variable which is obtained through consultation with concerning engineering teams. The provided opinions by the experts that are fuzzy labels of possibility are FIS rules which are stored in the rule base block whose formation is depicted at the bottom right-hand side of Figure 3. The two initial steps for rule base block formation are discussed in 5.4 and 5.6.

5.6. Questionnaire partitioning

FIS development wasn't possible if all the level combinations of the input variables didn't exist in the questionnaire and this caused too many cells which make human comparison quite inaccurate. In order to decrease the number of the comparisons and also to provide a guideline for the experts to help increase the accuracy, a zoning system is used based on how much stress a combination of contributing factors' levels (one of the questionnaire cells) creates for a component. A component is more likely to fail in a condition with a higher degree of stress. Thereby 5 stress levels were specified to create 5 regions (stress row in Table 2) for comparison instead of 27-factor level combinations (i.e. $3 \times 3 \times 3$). Experts were to fill out these regions by a set of fuzzy labels that was suggested for each stress region but they were free to choose other values for different combinations in the same stress region if they saw fit. Table 3 shows each stress level with its proper set of fuzzy labels. Table 4 shows membership function boundaries for the output variable.

Data was gathered from a group of 12 engineers with relative knowledge and enough experience from departments of management, maintenance, and design. Since opinions vary, aggregation is needed so that a single value is produced for each cell. To this end, a weighing factor was calculated for each engineer according to a weighing system in Table 5 so that a weighted mean could be calculated for the opinions. Table 6 shows the computed weighting factors for each engineer. The results of the weighted mean for the RV1 are presented in Table 2 as an example from which the rules for this component were extracted and were written below the Table.

5.7. Fuzzy Inference Fault Tree Analysis; FIFTA

At this point, FT is drawn, opinions are gathered to form the rule base block and membership functions are stored in the database block;

Table 3. Represented labels by stress regions

| stress regions | code | represented levels | Fuzzy labels |
|----------------|------|--------------------|--------------|
| Green | G | Very low | (1,5) |
| Blue | B | Low | (6,10) |
| Yellow | Y | Medium | (11,15) |
| Orange | O | High | (16,20) |
| Red | R | Very high | (21,25) |

Table 4. Fuzzy ranges of fuzzy labels

| label | Fuzzy range | label | Fuzzy range | label | Fuzzy range | label | Fuzzy range | label | Fuzzy range |
|-------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|-------------|
| 1 | (0,4,8) | 6 | (20,24,28) | 11 | (40,44,48) | 16 | (60,64,68) | 21 | (80,84,88) |
| 2 | (4,8,12) | 7 | (24,28,32) | 12 | (44,48,52) | 17 | (64,68,72) | 22 | (84,88,92) |
| 3 | (8,12,16) | 8 | (28,32,36) | 13 | (48,52,56) | 18 | (68,72,76) | 23 | (88,92,96) |
| 4 | (12,16,20) | 9 | (32,36,40) | 14 | (52,56,60) | 19 | (74,76,80) | 24 | (92,96,100) |
| 5 | (16,20,24) | 10 | (36,40,44) | 15 | (56,60,64) | 20 | (76,80,84) | 25 | (96, 100) |

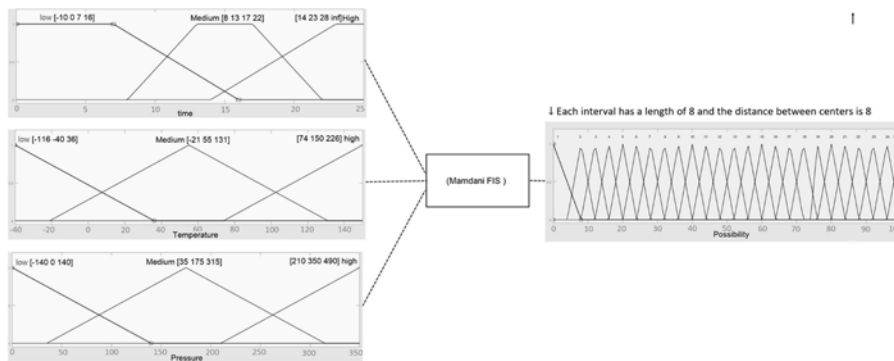


Fig. 7. Generation of a single output in FIFTA for an FTA with (j) basic events

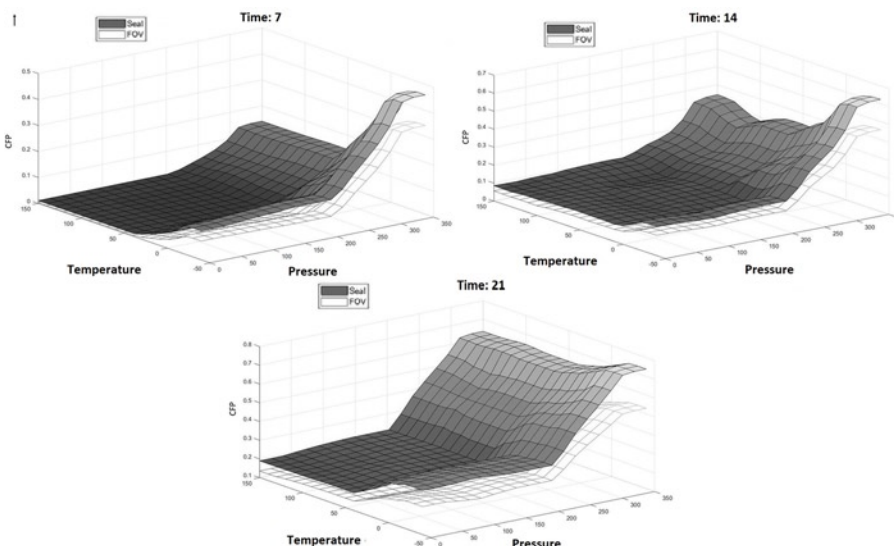


Fig. 8. Cumulative failure probability versus different operational conditions after running system for 7, 14, and 21 years lifetime

So, an FIS box is formed for each basic event which leads to the point where FISs should be linked to the drawn FT.

The desired combination of variable scalar ranges (e.g. 7th year in a pressure of 150 bars and a temperature of 50°C) is selected as an op-

erating condition (i.e. combination (i)). This combination is inserted as an input for each FIS and a probability value is generated for each basic event. Basic event probability values are inserted into the FTA formula and the probability of the top-event is calculated (i.e. probability (i)). Then a new combination is selected to generate a new top-event probability and so on; until enough points are generated for the response surface to be drawn. The result is a n-dimensional surface (n-1: the number of contributing factors) whose vertical axis show FT output or cumulative failure probability (CFP) of the proposed alternative. The horizontal axes show the contributing factors. This process is presented in the middle of Figure 3. Figure 7 illustrates a more detailed description of the process of generating one output point.

The response on a 3-D surface is drawn for both systems, presented in Figure 8. Time is separated as the 4th dimension. The influence of time and the two other contributing factors can be seen simultaneously which is the unique trait of this technique.

6. Discrete-event simulation

In this section, the FTA method will be used again as the core of the method, and in addition, discrete event simulation will be used to evaluate the system reliability.

One of the applications of discrete event simulation is in assembly and production systems and the use of this tool develops this capability for managers and engineers to gain a broad understanding of their system and can evaluate the effect of a small change in the whole system. And thus be able to calculate the reliability of the system. For example, suppose that by making a change in a station in the system, we have caused changes in the performance of that station. These changes may be predictable because the system under study is extremely small and its relationship with other components has not been studied. But answering the question of what effect the changes made in this station will have on the efficiency and reliability of the whole system and on other stations is a question that is very difficult to answer without using simulation tools. In many cases it is impossible. In this regard, in this section, a discrete-event simulation is implemented to evaluate and compare the reliability of two common flare systems.

6.1. Gathering input data

The input data actually provides the driving force for the simulation model. The steps that need to be taken to create an efficient model for the input data are:

- a) data collection,

Table 5. Scoring system

| Constitution | Classification | Score |
|-----------------------|---|-------|
| Professional position | Professor, GM/DGM, chief Engineer, Director | 5 |
| | Asst. prof, Manager, Factory inspector | 4 |
| | Engineer, supervisor | 3 |
| | Foreman, technician, graduate | 2 |
| | Apprentice operator | 1 |
| Service time | >30 years | 5 |
| | 20-30 | 4 |
| | 10-20 | 3 |
| | 5-10 | 2 |
| | <5 | 1 |
| Education level | Ph.D./M.Tech. | 5 |
| | M.Sc./B.Tech. | 4 |
| | Diploma/B.Sc. | 3 |
| | ITI | 2 |
| | technical college | 1 |
| Age | >50 | 5 |
| | 40-50 | 4 |
| | 30-40 | 3 |
| | 25-30 | 2 |
| | <25 | 1 |

- b) selecting the input probability distribution and determining the parameters of the selected probability distribution,
c) evaluating the selected distribution and its related parameters for the goodness of fit.

In order to collect data, the following methods were used:

- observing the system and collect sufficient samples of each process,

- interviews with related experts,
- imaging, video recording, and recording of system processes,
- collecting raw data from software available at the refinery.

After collecting the required data, random variables were modeled using a candidate probability distribution. Hence any statistical package may be applied. Table 7 prepared a list of the best probability fitting function as well their relevant estimated parameters.

6.2. Simulation model

After collecting all the necessary information from each of the flare gas recovery systems, and fitting the appropriate distributions for the data, a computer simulation of the systems was performed. In this research, the Arena software has been used for simulation. Arena is an application software for simulating discrete event systems. Arena is complete software for simulation studies and supports all steps of a simulation study. Arena provides templates that make it easy to create the right animation for simulation issues. Templates are a group of modules that contain entities, processes, and special language for a specific type of problem. Arena has an input analyzer and an output analyzer. The user can view the raw data input using the analyzer. The output analyzer is also for viewing and analyzing simulation data.

The settings of the simulation model components are mentioned as below.

A) Observation Period:

Since the work schedule of the flare gas system is usually determined at the beginning of each month, the observation period of each simulation sub run is considered to be 30 working days.

B) Number of replications:

In order to achieve acceptable results and reduce the length of the confidence interval of system performance criteria, it is necessary to run a simulation model for a significant number of replications. The number of replications of the simulation is determined according to the half-width of the system performance criteria. The most important performance measure for this purpose is the average system reliability. Our experiments showed that if we consider the number of replications of the simulation as 90, the half-width of the above performance criteria has reached an acceptable level and is about 1 to 3% of the average.

Table 6. Experts' scores and calculated weighting factors

| # Expert | Title | Score | Service time (years) | Score | Education level | Score | Age | Score | Weighting score | Weighting factor |
|----------|---|-------|----------------------|-------|-------------------|-------|-------|-------|-----------------|------------------|
| 1 | Engineer, supervisor | 3 | 20-30 | 4 | ITI | 2 | 25-30 | 2 | 11 | 0.09 |
| 2 | Apprentice operator | 1 | <5 | 1 | technical college | 1 | <25 | 1 | 4 | 0.03 |
| 3 | Foreman, technician, graduate | 2 | 5-10 | 2 | M.Sc./B.Tech. | 4 | <25 | 1 | 9 | 0.08 |
| 4 | Foreman, technician, graduate | 2 | 10-20 | 3 | Diploma/B.Sc. | 3 | <25 | 1 | 9 | 0.08 |
| 5 | apprentice operator | 1 | 10-20 | 3 | technical college | 1 | 25-30 | 2 | 7 | 0.06 |
| 6 | Engineer, supervisor | 3 | <5 | 1 | M.Sc./B.Tech. | 4 | >50 | 5 | 13 | 0.11 |
| 7 | Asst. prof, Manager, Factory inspector | 4 | 5-10 | 2 | ITI | 2 | 25-30 | 2 | 10 | 0.08 |
| 8 | Professor, GM/DGM, chief Engineer, Director | 5 | <5 | 1 | ITI | 2 | <25 | 1 | 9 | 0.08 |
| 9 | Engineer, supervisor | 3 | 10-20 | 3 | M.Sc./B.Tech. | 4 | <25 | 1 | 11 | 0.09 |
| 10 | Engineer, supervisor | 3 | >30 | 5 | technical college | 1 | 25-30 | 2 | 11 | 0.09 |
| 11 | Apprentice operator | 1 | 5-10 | 2 | M.Sc./B.Tech. | 4 | 25-30 | 2 | 9 | 0.08 |
| 12 | Professor, GM/DGM, chief Engineer, Director | 5 | >30 | 5 | ITI | 2 | >50 | 5 | 17 | 0.14 |
| | | | | | | | | | | |
| | Total sum | | | | | | | | 120 | 1.00 |

C) Warm-up Period:

In order for the simulation model to reach a steady-state and the output of the model to be calculated in a steady state, a warm-up time is mainly considered for the system. This time period only plays the role of warming up and stabilizing the system performance criteria and has no role in the final calculations.

In order to calculate the system warm-up time period, the behavior of some system performance criteria has been examined and the time it takes for them to reach a steady-state has been considered as the system warm-up time period.

Figure 9 shows the trend chart of the average system reliability in three different replications. As can be seen from Figure 9, in all replications after a period of 4 days, the reliability of the system has reached a stable state. Therefore, the warm-up period of the system is 4 days.

D) Verification of the Simulation Model:

One of the basic steps after creating a simulation model is to check the verification of the model. In this section, it should be checked whether the structure of the simulation model is based on the conceptual model and its hypotheses. There are different methods to check the verification of the model. In this study, the following steps were performed to verify the model:

- Checking software sub-models and debugging software codes.
- A more detailed review of the model by other experts.
- Checking model outputs for different inputs.
- Checking the model step by step and compare the output of mode variables with manual calculations.
- Preparation of two-dimensional and three-dimensional animation of the model to understand and correct mistakes.

E) Validation of the Simulation Model:

Validation is the study of whether the conceptual model and the specific model created accurately represent the system under study. Since simulation is an estimate of the real world, it should be noted that it is not possible to validate 100% of the model with the real system. In this research, the three-step method proposed by Naylor and Finger has been used:

Step 1: To develop a model with high frequent validity

The purpose of the first stage is to create a model that has the most apparent validity so that it seems logical from the point of view of the people in the model system. In this section, sensitivity analysis was used to check the apparent validity of the model; in this way, we changed the failure rate of system components and examined its impact on system reliability. It is clear that as the failure rate decreases, the reliability of the system must increase.

Step 2: An Empirical Investigation of model hypotheses

In this step, two main categories of model hypotheses related to model structure and related to model information were examined. The above hypotheses were tested experimentally and intuitively with the cooperation of refinery experts.

Step 3: Examining the simulation outputs

The most effective consideration for validating the model is that the simulation outputs should not be as significantly different as possible from the actual process outputs. For this purpose, the hypothesis test method has been used to validate the model outputs. In this study, the amount of system exhaust gas in has been selected as a criterion for comparison with the real system and validation of simulation outputs. Here, the unit of measurement of gas exhaust is reported by MSCMD (Million Standard Cubic Meter per Day). Each cubic meter per day (m³/d) of flow rate equals: 0.000035 million standard cu-ft of gas per day (at 15°C).

In order to validate the model, the average exhaust gas of the simulation model (Y_1) was compared with the actual system average (Z_1) and the following hypothesis was tested:

$$H_0 : E(Y_1) = Z_1 = 65000 \quad H_1 : E(Y_1) \neq Z_1 = 65000$$

If the H_0 hypothesis is not rejected, then there is no reason to reject the equality of the model exhaust gas averages and the actual system exhaust gas. If the assumption H_0 is rejected, then the assumption of the equality of the means of the exhaust gas of the model and the actual exhaust gas of the system is rejected and the model is not valid.

The results of the hypothesis test at a significance level of $\alpha = 0.05$ are as follows:

- Test of $\mu = 65000$ vs not = 65000,
- $N = 30$,
- Mean = 66342,
- Standard Deviation = 1065,
- 95% Confidence interval = (62436, 67596),
- P-value = 0.067.

Since the P-value (0.067) is greater than the significance level (0.05), there is no reason to reject the H_0 hypothesis. Looking at the results of the above hypothesis test, we find that there is no significant difference in 95% confidence level between the outputs of the simulation model and the outputs of the real system; Therefore, the resulting simulation model is valid.

7. Results and discussion

Using FIFTA, a sufficient number of points are generated to draw a surface for each alternative. The cumulative probability of failure surface is a functionality associating an operating condition to a probability value. This is the required function described in the problem statement that associates a point in its domain (i.e. the space created by axes of time, pressure, and temperature and limited to their boundaries) to a value in its codomain (i.e. CFP); which demonstrates each alternative's resilience under different operating conditions. It should be pointed out that cumulative failure probability is drawn instead of reliability to have a convex function for a better illustration. It is known that $R(t) = 1 - F(t)$ so a rise in CFP means a fall in reliability.

In order to illustrate the surfaces, one of the dimensions of the domain space needs to be separated so that the surface is drawn on a plane. The "Time" axis is separated to study the changes of reliability on the pressure-temperature plane. This provides the opportunity for the decision-makers to investigate the effects of the behaviors of gas on the system's failure probability in its different age periods.

Since FIFTA is being used to generate data, the surfaces on the pressure-temperature plane can be drawn for any age period of the system. 25 surfaces were drawn for the systems for each year, and the surfaces with the most significant changes were chosen to be illustrated in this paper.

The CFP surfaces were drawn so that reliability differences would help make a choice between the proposed alternatives. In the presented graphs, the CFP surface of the FOVS is always below that of the SDS meaning that the reliability of FOVS is higher than that of SDS in all operating conditions. Thus, FOVS outperforms SDS for concerned refineries and could be installed prior to the flaring tower.

In order to have a simplified representation of the drawn graphs, the pressure-temperature plane is divided into nine areas, seen in Table 8.

Each of these areas stands for a general operating condition where a system has relatively similar behavior. A proper number of points on each surface are selected in each area and an average of their CFPs (ACFP) is calculated. The results can be seen in tables (9, 10, 11). The last column of the Tables shows the percentage of difference of the

Table 8. NUMBER OF Divided areas on the temperature, pressure plain

| Press temp | [0,200] | [200,300] | [300,350] |
|------------|---------|-----------|-----------|
| [100, 150] | 1 | 4 | 7 |
| [0, 100] | 2 | 5 | 8 |
| [-50, 0] | 3 | 6 | 9 |

ACFPs (as a representative of reliability and performance) between SDS and FOVS. As expected, there is always a positive difference in the last column because the CFP surface of the SDS is always above that of the FOVS.

Besides the results obtained for the alternatives studied in this paper, in other cases after drawing the surfaces, there might not be a clear winner. In some cases, the surface of an alternative may be partially above and partially below that of the other alternative, which shows different resilience in different operating conditions. Thus, in the above Tables, some of the calculated numbers in the last columns would be negative. This could make the decision-making a lot more complicated.

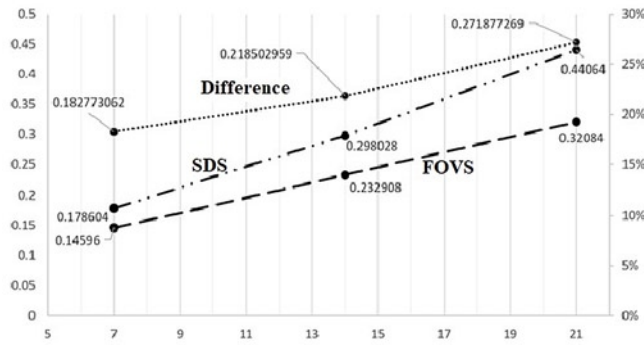


Fig. 9. The warm-up period in the simulation model

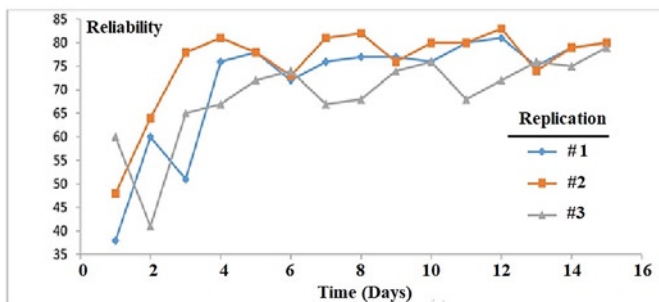


Fig. 10. Weighted average of ACFPs versus time (year)

To make a decision between such alternatives, different scores could be attributed to each of the 9 areas. This way, equal reliability values in different operating conditions would not be equally significant. Scores for each area can be based on a number of factors (e.g., the percentage of time they occur, how costly the type of the damage caused by an operating condition can be, the likelihood of failure of the systems in each area, etc.). It is up to the decision-making team to differentiate the importance of good performance in an area. The attributed scores by the decision-making team, are seen in column 2 of the above-mentioned Tables.

The score of an area can be used as a weight for the ACFP of that area to calculate a weighted average and have a single numerical value for the whole surface in a time period. Based on the calculated values for each year, a 2D graph is drawn in Figure 10 to show the difference in the performance of the alternatives versus time.

References

1. Attar Ahmad, Raissi Sadigh, Khalili-Damghani Kaveh. A simulation-based optimization approach for free distributed repairable multi-state availability-redundancy allocation problems, *Reliability Engineering & System Safety*. 2017; 157: 177-191, <https://doi.org/10.1016/j.res.2016.09.006>.
2. Azadeh A, Ebrahimipour V, Bavar P. A fuzzy inference system for pump failure diagnosis to improve maintenance process: The case of a petrochemical industry. *Expert Systems with Applications* 2010; 37(1): 627-639, <https://doi.org/10.1016/j.eswa.2009.06.018>.
3. Baig AA, Ruzli R. Estimation of failure probability using fault tree analysis and fuzzy logic for CO2 transmission. *International Journal of Environmental Science and Development* 2014; 5(1): 26.

Table 12 shows the weighted average of ACFPs. As seen in Figure 10, there is a clear advantage to using FOVS since it has a lower cumulative failure probability during its life (22.4% difference in average). Also, it can be seen that the difference in performance gets larger with the passage of time which concludes the comparisons.

Individual assessments can also be made on each alternative using the surfaces, and the following information might be of interest for the design team, maintenance team, and the management: The most dangerous scenario that can happen for the safety system is when gas passes through the systems with high pressure (250, 350) and low temperature (-50, 0) where the likelihood of failure is at its maximum level. Besides that, the safest operating condition is now detected in Figure 8.c where systems are in a high age. A combination of pressure of (50, 200) and a temperature of (50, 150) is the safest operating condition where both systems have the highest reliability level. The minimum level of pressure is considered "50" for a minimum flow that avoids flashbacks.

The above paragraph highlights another possible usage for the results obtained from FIFTA in cases where the contributing factors can be brought under control. Using the resulting surfaces from FIFTA, one can identify the best operating areas where reliability value is higher and keep the levels of the contributing factors in the identified areas. These are the standard limits that can be implemented in monitoring or controlling subsystems.

Other than that design improvements can be made in a system by identifying the operating conditions causing the lowest reliabilities. Then, if possible, sensitive components to that operating condition can be replaced with the ones that are more resilient against them (e.g. if high temperature decreases the reliability, high-temperature resilient components that can be used in the system).

8. Conclusion

In the current research, we showed that it is possible to obtain an interactive output result from FTA by fusing FIS and discrete-event simulation so that output changes can be identified for different contributing factors. The proposed expert-based approach and zoning system can help gather the required information for calculations in the purchasing phase. This provides a practical approach towards prognostic studies when actual assessments haven't been performed on a system.

We showed that from the two proposed alternatives as a safety sub-system for an FGRU, FOVS outperforms SDS in a different age in terms of reliability judging by the lower CFP, and since the systems are assessed in different operating conditions, the comparison is fairly comprehensive which makes the final decision highly justified. Taking multiple factors into account helps also prevent the unforeseen failures of the safety subsystem.

The generated surfaces can also provide insight for design enhancements and control processes by indicating the system's resilience towards different operating conditions. Also, if there is the possibility to control the contributing factors, the surfaces can provide an approximation of standard limits for their levels. Here, judging by the generated results, it is suggested that the winning alternative isn't exposed to a simultaneous rise in gas pressure and temperature due to the massive plunge of reliability in this area.

4. Briš R, Praaks P. Simulation approach for modeling of dynamic reliability using time dependent acyclic graph. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2006; (2): 26-38.
5. Choi IH, Chang D. Reliability and availability assessment of seabed storage tanks using fault tree analysis. *Ocean Engineering* 2016; 120:1-14, <https://doi.org/10.1016/j.oceaneng.2016.04.021>.
6. Elsayed T. Fuzzy inference system for the risk assessment of liquefied natural gas carriers during loading/offloading at terminals. *Applied Ocean Research* 2009; 31(3): 179-185, <https://doi.org/10.1016/j.apor.2009.08.004>.
7. Elvidge CD, Bazilian MD, Zhizhin M, Ghosh T, Baugh K, Hsu F-C. The potential role of natural gas flaring in meeting greenhouse gas mitigation targets. *Energy Strategy Reviews* 2018; 20:156-162, <https://doi.org/10.1016/j.esr.2017.12.012>.
8. Khodayee SM, Chiacchio F, Papadopoulos Y. A Novel Approach Based on Stochastic Hybrid Fault Tree to Compare Alternative Flare Gas Recovery Systems. *IEEE Access* 2021; 9: 51029-49, <https://doi.org/10.1109/ACCESS.2021.3069807>.
9. Nematkhah Farnam, Raissi Sadigh, Ghezavati Vahidreza. An Integrated Fuzzy DEMATEL-Fuzzy ANP Approach to Nominate Diagnostic Method and Measuring Total Predictive Performance Score, Safety and Reliability 2017; 37(1): 37-42, <https://doi.org/10.1080/09617353.2017.1411676>.
10. Pietrusiak D. Evaluation of large-scale load-carrying structures of machines with the application of the dynamic effects factor. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2017; 19(4) :542-551, <https://doi.org/10.17531/ein.2017.4.7>.
11. Pourhassan MR, Raissi S, Apornak A. Modeling multi-state system reliability analysis in a power station under fatal and nonfatal shocks: a simulation approach. *International Journal of Quality & Reliability Management* 2021, <https://doi.org/10.1108/IJQRM-07-2020-0244>.
12. Pourhassan Mohammad Reza, Raissi Sadigh, Hafezalkotob Ashkan. A simulation approach on reliability assessment of complex system subject to stochastic degradation and random shock. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2020; 22 (2): 370-379, <https://doi.org/10.17531/ein.2020.2.20>.
13. Purba JH. A fuzzy-based reliability approach to evaluate basic events of fault tree analysis for nuclear power plant probabilistic safety assessment. *Annals of Nuclear Energy* 2014; 70: 21-9, <https://doi.org/10.1016/j.anucene.2014.02.022>.
14. Raissi Sadigh, Ebadi Shiva. A Computer Simulation Model for Reliability Estimation of a Complex System. *International Journal of Research in Industrial Engineering* 2018; 7(1): 19-31, <https://doi.org/10.22105/RIEJ.2018.109017.1032>.
15. Rajakarunakaran S, Kumar AM, Prabhu VA. Applications of fuzzy faulty tree analysis and expert elicitation for evaluation of risks in LPG refueling station. *Journal of Loss Prevention in the Process Industries* 2015; 33: 109-23, <https://doi.org/10.1016/j.jlp.2014.11.016>.
16. Ratnayake Chandima R.M., Application of a fuzzy inference system for functional failure risk rank estimation: RBM of rotating equipment and instrumentation. *Journal of Loss Prevention in the Process Industries*. 2014; 29: 216-224, <https://doi.org/10.1016/j.jlp.2014.03.002>.
17. Renjith V, Madhu G, Nayagam VLG, Bhasi A. Two-dimensional fuzzy fault tree analysis for chlorine release from a chlor-alkali industry using expert elicitation. *Journal of Hazardous Materials* 2010; 183(1-3): 103-110, <https://doi.org/10.1016/j.jhazmat.2010.06.116>.
18. Saeidi RG, Amin GR, Raissi S, Gattoufi S. An efficient DEA method for ranking woven fabric defects in textile manufacturing. *International Journal of Advanced Manufacturing Technology* 2013; 68 (1-4): 349-354, <https://doi.org/10.1007/s00170-013-4732-4>.
19. Selech J, Andrzejczak K. An aggregate criterion for selecting a distribution for times to failure of components of rail vehicles. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2020; 22 (1): 102-111, <https://doi.org/10.17531/ein.2020.1.12>.
20. Wang D, Zhang P, Chen L. Fuzzy fault tree analysis for fire and explosion of crude oil tanks. *Journal of Loss Prevention in the Process Industries* 2013; 26(6): 1390-1398, <https://doi.org/10.1016/j.jlp.2013.08.022>.
21. Yazdani E, Asadi J, Dehaghani YH, Kazempoor P. Flare gas recovery by liquid ring compressors-system design and simulation. *Journal of Natural Gas Science and Engineering* 2020; 84: 103627, <https://doi.org/10.1016/j.jngse.2020.103627>.
22. Zhang X, Sun L, Sun H, Guo Q, Bai X. Floating offshore wind turbine reliability analysis based on system grading and dynamic FTA. *Journal of Wind Engineering and Industrial Aerodynamics* 2016; 154: 21-33, <https://doi.org/10.1016/j.jweia.2016.04.005>.

Use of emission indicators related to CO₂ emissions in the ecological assessment of an agricultural tractor

Łukasz^a Rymaniak, Jerzy Merksiz^a, Natalia Szymlet^a, Michalina Kamińska^a, Sylwester Weymann^b

^aPoznań University of Technology, Faculty of Civil and Transport Engineering, Institute of Combustion Engines and Powertrains, ul. Piotrowo 3, 61-138 Poznań, Poland

^bŁukasiewicz Research Network - Industrial Institute of Agricultural Engineering, ul. Starołęcka 31, 60-963 Poznań, Poland

Indexed by:



Highlights


- The dimensionlessness of the emission indicator M, it is possible to compare not only vehicles of the same category, but also of objects of different purposes.
- It is necessary to determine the resistances of the internal combustion engine and take into account their values when determining the torque in tests carried out during the vehicle operation and the work performed by the drive system.
- Increasing the driving speed of the tractor during typical field work from 5 km/h to 15 km/h may have a positive impact on the overall exhaust emission results of toxic compounds.

Abstract

The paper presents the proposed proprietary M exhaust emission indicator, which is based on the assumption that CO₂ emissions are a measure of the correctness of the combustion process. The measurements were performed using a farm tractor meeting the Tier 3 emission norm, operated in real conditions during plowing work. The tests were carried out for a given land section at three speeds. In the analysis of test results, the net engine work was used, as it is carried out in the type approval procedures. When measuring in real operating conditions, the torque read from the OBD system is overstated because it takes into account the engine's internal resistance. In the analysis of test results, the fuel consumption, emission indicators of gaseous compounds and particulates were determined, and the best conditions for conducting agricultural works were indicated in terms of their impact on the natural environment. The aim of the work is to verify the possibility of determining the emission index for an off-road vehicle and a comparative analysis of its values for various operating parameters of a farm tractor. On this basis, it was found that the lowest values of the M identity were recorded for the test characterized by a vehicle speed of 15 km/h.

Keywords

agricultural machinery, emission of CO₂, emission indicators, PEMS, RDE.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

1. Introduction

Manufacturers of agricultural vehicles and machines equipped with internal combustion engines make continuous efforts to reduce the negative impact of their products on the natural environment. It is required both by the legislative guidelines defined in given areas of the world, as well as by the increasing ecological awareness of the human population [2]. The main disadvantages of modern internal combustion engines include noise emission and emission of toxic compounds into the atmosphere. It should be noted, however, that in relation to combustion engines produced even ten years ago, the emission factors decreased by at least several dozen percent. In parallel, considerations are being made to introduce a reduction in CO₂ emissions, which is equivalent to a reduction in fuel consumption.

Agricultural motor vehicles in different regions of the world/countries have to undergo a number of different type approval procedures, including those related to the exhaust emission of pollutants. In the European Union, non-road vehicles must meet the Stage standards, which categorize the engines in terms of their intended use, power and emission indicators as shown, among others, by in the latest regulation of the European Union [20]. Until now, legislative tests have been performed only for the internal combustion engines themselves on

engine laboratory dynamometers. These engines, depending on their intended use, are most often tested in two basic types of tests: static (including the *Non-Road Stationary Cycle* NRSC) and dynamic (including the *Non-Road Transient Cycle* NRTC). Currently, since 2019, guidelines for the monitoring of exhaust emissions of gaseous components during real operation also begin to apply for selected NRMM subgroups, but the final limits are still not fully defined.

In the last 20 years (since 1997), the European Commission has presented 7 Directives which include type approval guidelines for off-road vehicles in terms of exhaust emissions. In 2016, Regulation (EU) 2016/1628 was introduced, which presented new exhaust emission limits for off-road vehicles in the Stage V norm, which has been in force since 2019. This document is the same for all EU Member States, as so far specific additional guidance (e.g. on particle number limits) existed only in some countries. The new document also covers a wider range of combustion engines: less than 19 kW and over 560 kW in power. Figure 1 shows Changes in HC + NO_x and PM emission limits for Stage I-V standards for an exemplary group of combustion engines of off-road machines. The presented relationships indicate that the PM limit in the Stage V standard is 97% lower than in the Stage I, and the HC + NO_x limit has been reduced by 94%.

E-mail addresses: Ł. Rymaniak - lukasz.rymaniak@put.poznan.pl, J. Merksiz - jerzy.merksiz@put.poznan.pl, N. Szymlet - natalia.r.szymlet@doctorate.put.poznan.pl, M. Kamińska - michalina.kaminska@put.poznan.pl, S. Weymann - weymann@pimr.poznan.pl

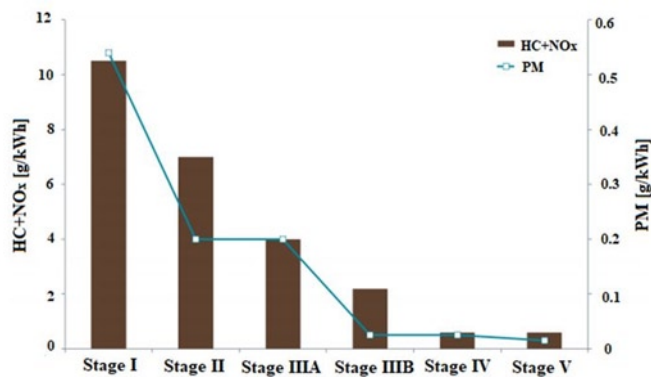


Fig. 1. Exhaust emission limit changes of HC+NO_x and PM for the Stage I–V norms [11]

The development and miniaturization of exhaust measuring equipment belonging to the Portable Emissions Measurement System (PEMS) group, which has been progressing in recent years, allows for increasingly more precise tests of the environmental performance of vehicles in real operating conditions to be performed. Currently, these types of testing and research activities are conducted all over the world [1, 3, 4]. They are necessary because, as the studies [12–14] prove, qualitative and quantitative exhaust emission measurements in type-approval tests and in actual operation differ significantly. Road conditions are characterized by an unfavourable effect, primarily on PM and NO_x emissions [9, 10, 21]. Therefore, with Stage V, tests in real operating conditions were to come into force, which have not yet been defined. In terms of particulate emissions, the current limits apply to their mass. Along with the legislative changes, however, limits for the number of particulates are being introduced for the engines of non-road mobile machinery of the NRE category (*Engines for Non-road Mobile Machinery*), as in the case of other passenger and heavy vehicles.

2. Definition of agricultural vehicle emission indicators in relation to CO₂

Agricultural vehicles classified as non-road vehicles are mostly used off public roads in non-urbanized areas. However, due to their overall number, their direct impact on green areas and agricultural crops is significant, as well as on the natural environment. Contemporary agricultural machinery is the result of great technological progress. Currently, their construction uses advanced technologies, safety and comfort systems and non-motor systems that require the use of extensive electronics. This increases the cost of producing the product, but it undoubtedly has a very positive impact on the ecological indicators they currently achieve. Precise control of the fuel supply process enables more efficient use of fuel than in structures based solely on mechanical solutions. This is justified by the fact that in the considered group of machines, despite the increasingly more numerous equipment related to the driver's comfort and greater work efficiency, no significant increase in fuel consumption was observed. All over the world, activities are undertaken by leading scientific and research centers to assess the environmental performance of machines in real operating conditions – in the field. This allows for a broader understanding of the problem of emissivity and allows for the development of new solutions or modification of the existing ones in such a way as to minimize the negative impact of this group of machines on the natural environment as much as possible.

The process of burning fuel in an internal combustion engine is used to generate thermal energy. During its implementation, a number of harmful and toxic chemical compounds are formed. Carbon dioxide CO₂ is produced through complete oxidation. Its emission is not restricted by the current legislative guidelines in terms of a specific vehicle. Existing CO₂ limits in the European Union apply to car manufacturers, but these guidelines take into account the entire range of ve-

hicles produced. The CO₂ emissions of 95 g/km are the average value of the entire model range for each brand. These guidelines were to apply from 2020, but eventually did not enter into force [19]. These types of restrictions apply to passenger cars. Similar ones are introduced for heavy vehicles.

Carbon dioxide is not defined as a toxic exhaust substance, but only as harmful one. It is the main cause of the greenhouse effect and in higher concentrations is poisonous to living organisms. Carbon monoxide CO, hydrocarbons THC and particulates PM (in terms of mass and number) are all formed in the combustion process during partial and incomplete oxidation, while nitrogen oxides NO_x form in the presence of high temperatures. The mentioned combustion products adversely affect the natural environment and it is necessary to introduce solutions limiting their concentration during vehicle operation. They are all undesirable products that significantly contribute to the environmental degradation and pollution.

For working machines, in research works, special units are proposed that refer to the mass of the emitted exhaust component in relation to the work performed (e.g. area of plowed ground, volume of felled tree, etc.). The emission factors developed in this way are sometimes difficult to compare between vehicles (machines), which depends, inter alia, on their intended use, type of drive system used (e.g. hybrid), operating conditions, driving style, etc. For this reason, a new proprietary index was proposed, based on the use of measurements of carbon dioxide parameters (e.g. the emission intensity of this compound, its road or unit emission). The values that are substituted into its structure must be expressed in the same units, which will make it dimensionless.

The physical and chemical processes in the cylinder of an internal combustion engine are complex and very difficult to fully define or simulate. Taking into account the basic combustion equations, it can be assumed that the ratio of CO₂ to the remaining toxic components of the exhaust gas is a measure of the correctness and efficiency of the fuel combustion process. Comparing the emission of toxic compounds with the emission of carbon dioxide, it was proposed to determine the proprietary emission factors M, which characterize a given combustion unit or power unit (if the system will also test non-engine exhaust gas cleaning systems). Such defined ecological indicators allow for the effective comparison of various heat engines with exhaust gas treatment systems. From this perspective, it is a new way to assess the environmental performance of a given vehicle / machine. To achieve this, it is necessary to use the quantitative dimensionless quantitative emission factor M, which is defined by the quotient:

$$M_j = b \cdot \frac{e_{rzecz,j}}{e_{CO_2}} \quad (1)$$

where: M – dimensionless emission indicator [–]; j – toxic exhaust component for which the emission indicator was determined; b – universal constant (for CO, THC and NO_x = 10³, for PM = 10⁵); e_{rzecz,j} – specific emission, road emission or mass of toxic compound j determined during the measurements in the emission test [g/(kW·h); g/(km); g]; e_{CO₂} – specific emission, road emission or mass of CO₂ determined during the measurements in the emission test (having the same unit as e_{rzecz,j}) [g/(kW·h); g/(km); g].

The presence of the constant b allows to increase the readability of the results, because the number of decimal places is limited. This has been confirmed on other vehicles that meet various emission standards [16, 17]. Taking into account the internal combustion engine with non-engine exhaust aftertreatment systems, it is possible to consider the environmental impact of vehicles of various categories, especially in the road or field tests. Due to the dimensionless nature of the used indicator, it is also a good method for defining the exhaust emission of toxic compounds in relation to fuel consumption. Therefore, it is possible to ecologically assess the agricultural tractor (which is the subject of this study) and vehicles of

other categories based on the proposed M factor in terms of emission tests obtained both in laboratory tests and in real operating conditions. The proprietary M emission index was presented at the Real Driving Emission conference in Berlin in 2017. The idea was approved by the scientific world and representatives of the European Commission. This confirms that the inclusion of the M index in the assessment of the environmental performance of motor vehicles is justified and is characterized by innovation on a global scale. The research carried out so far has not presented the results of work on such defined issues. Thus, an extensive literature review showed that the only publications on this subject belong to the authors of this work [16, 17]. The dissemination of the indicator presented allows for even more precise definition of the environmental performance of the vehicle / machine.

3. Research method

The subject of the pollutant emissions tests in real operating conditions was an agricultural tractor belonging to the NRMM (*Non-Road Mobile Machinery*) vehicle group. The test vehicle was manufactured in 2007 and thus homologated according to the Stage IIIA/Tier3 norms. The manufacturer equipped the vehicle with a 6.7-liter compression-ignition engine, with a maximum power of 116 kW. The tractor was equipped with a DPF particulate filter, a DOC oxidizing catalytic converter and an EGR exhaust gas recirculation system. Before the tests were performed, the test object was inspected for possible technical defects or damage. In the field work, a cultivating unit was used, which loosened the soil and prepared it for sowing for cultivation. Measurements were made in a field in the town of Borek Wielkopolski. Figure 2 shows the test vehicle along with its technical specification.



| Parameter | Value |
|-----------------|--------------------------------------|
| Displacement: | 6,7 dm ³ |
| Driving gear: | 4-cylinder, electro-hydraulic clutch |
| Maximum power: | 116 kW/2100 rpm |
| Maximum torque: | 700 Nm/1250-1550 rpm |
| Equipment: | VGT Turbocharger, EGR, DOC, DPF |
| EURO standard: | Stage IIIA/Tier 3 |

Fig. 2. Picture of the tested vehicle and its technical data

The mobile measuring device SEMTECH DS was used in the research, which enables the performance of exhaust emission measurements in real operating conditions [6-8, 22]. It consists of a set of analyzers that allow determining the concentration of the basic gaseous exhaust gas components. The device was designed for use in measuring the exhaust emissions of machines/vehicles with compression ignition and spark ignition engines, compliant with the Stage II and newer emission norms. The device cooperates with the exhaust gas flowmeter, from which the exhaust gas sample is taken. It is transported by a heated line to the inside of the device. The following gaseous components are measured: THC (FID analyzer - *Flame Ionization Detector*), NO_x (NDUV analyzer - *Non-dispersive Detector Ultra Violet*), CO_x (NDIR analyzer - *Non-dispersive Detector Infra Red*), and the oxygen concentration is also measured (using

the electrochemical method) [13]. Measurement and data acquisition takes place at a frequency of 1 Hz. The error of the operation of the individual exhaust gas flowmeters and analyzers shall not exceed $\pm 3\%$. The measuring system also has a GPS device and a weather station (Fig. 3).

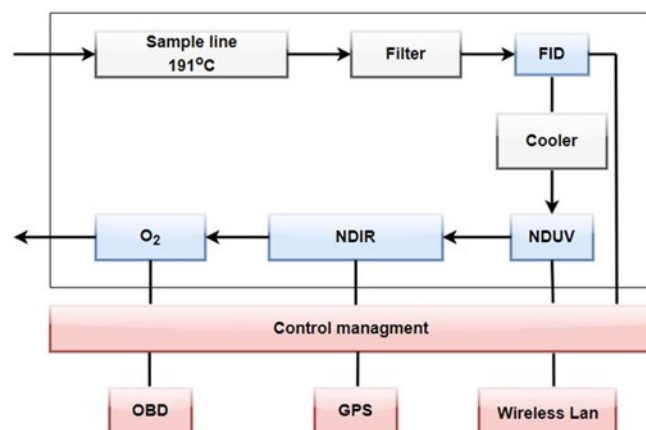


Fig. 3. Diagram of operation of PEMS- SEMTECH DS apparatus [23]

4. Test vehicle operating conditions

The tests of the agricultural tractor were carried out in real operating conditions, while working in the field. The measurements were made in three test cycles divided according to the speed during the work of the tested vehicle – done at 5, 10 and 15 km/h. Using the data recorded from the GPS, the operating parameters of the test vehicle were determined (Fig. 4). The share of operating time was presented relative to the variability of vehicle speed and acceleration. The characteristics include all research tests – for each speed including: acceleration, constant speed operation, and braking. For driving at constant speed (where $a = 0 \text{ m/s}^2$), the shares of 49%, 24%, and 15.4% were obtained, respectively, for the following value intervals: (1 m/s; 1.5 m/s), (2.5 m/s; 3 m/s) and (4 m/s; 4.5 m/s). In all tests, data was also acquired during stops occurring before and after the performed field work. The total share of time for this point was 3.2%. For all operating points including acceleration or braking (where $a \neq 0 \text{ m/s}^2$), their share was lower than 0.3%, which proves that the tests were performed correctly, i.e. that the assumed constant velocities were overall maintained during operation.

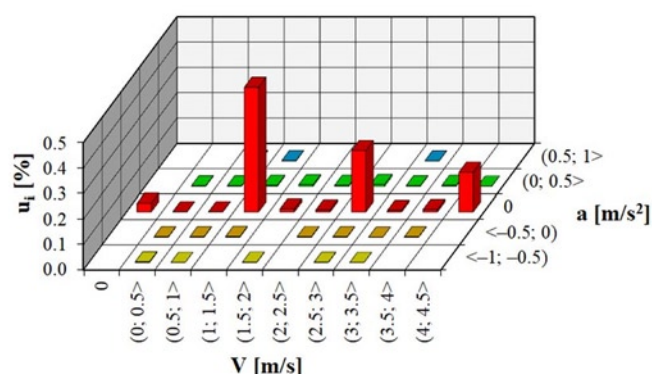


Fig. 4. The area of variability of work parameters of the research object during measurement tests

Based on the information recorded from the on-board diagnostic system, the parameters of the internal combustion engine were also acquired (Fig. 5). In order to determine the actual work performed by the internal combustion engine, when determining the specific exhaust emissions of measured components, it is necessary to take into account the net parameters: power and load, i.e. those obtained at the end of the crankshaft with power used by additional

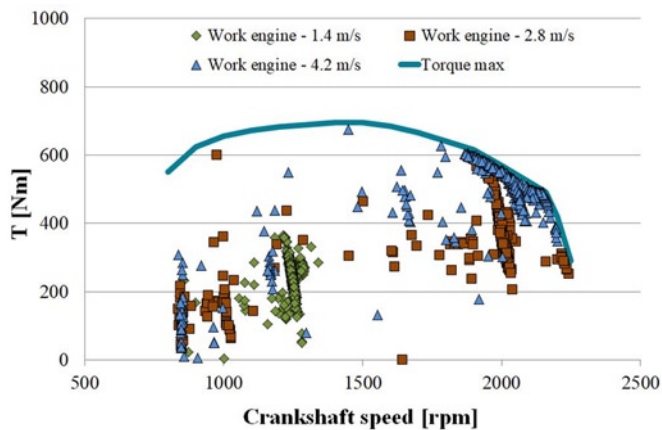


Fig. 5. Internal combustion engine operation parameters

devices already included. Crankshaft speed and generated torque values can be read from on-board diagnostic systems. The first of these parameters is determined directly using induction or Hall effect sensors, and the data obtained by this method is sufficient. However, the torque is determined on the basis of the pressure in the supply system and the injector opening time. However, there are some discrepancies in the real-world values as the readings obtained from the OBD system take into account the internal resistances of the engine. The calculations can address this problem by taking into account the percentage share of the load related to, among others, internal engine friction, however, this is a simplification as the actual internal resistances depend on many factors. As a rule, they are not linear and change depending on the current operating parameters of the engine. In selected areas of its operation, they may constitute up to 40% of the total torque produced. Its inflated value, which is read from the CAN (Controller Area Network), causes that the designated work in the test to also be overestimated. This translates into lowering the resulting specific exhaust emissions of pollutants, where the work performed by the drive system is in the denominator. For these reasons, it is necessary to determine the resistances of the internal combustion engine and take into account their values when determining the torque in tests carried out during the vehicle operation. This allows to determine the actual ecological indicators of a given vehicle [5].

For the first measurement test, the highest density of operating points occurred in the area of the load characteristic described by the engine speed interval (1150 rpm; 1245 rpm) and load interval (130 Nm; 358 Nm). During the second run, greater variability of operating parameters was observed. The distribution of work points also shows the greatest share of work in the load characteristic range, especially in the interval (1940 rpm; 2052 rpm) and (254 Nm; 560 Nm). While the highest engine speeds were achieved in the third test run, the engine was operated primarily at engine power band. This is evident for crankshaft speeds above 1,875 rpm. For all tests, there were fragments of operating time spent in the area of load characteristics that showed a lower density of operating points. This was the result of the internal combustion engine cooperation with the mechanical transmission and electro-hydraulic clutch used in the vehicle.

5. Results

In order to perform research on the evaluation of the emission indicators during a typical agrotechnical task performed by a farm tractor, the characteristics of the exhaust emission intensity as a function of engine rotational speed-load (n - T) were analyzed. The distribution of operating points covered the range of operating parameters in one-side closed intervals. The intensity of CO_2 emission increased with the increase of the engine load and rotational speed values (Fig. 6). The maximum CO_2 emission value (20.8 g/s) was recorded at a single operating point within the intervals of rotational speed (1800; 2000

rpm) and load (500; 600 Nm). The area where the highest intensity of carbon dioxide emission was recorded was observed for the ranges (1600; 2200 rpm) and (400–700 Nm). The average emission value for this interval was 19.07 g/s. The lowest emission intensity occurred in the load interval not exceeding 300 Nm. The mean value for the performed test drive was 11.1 g/s.

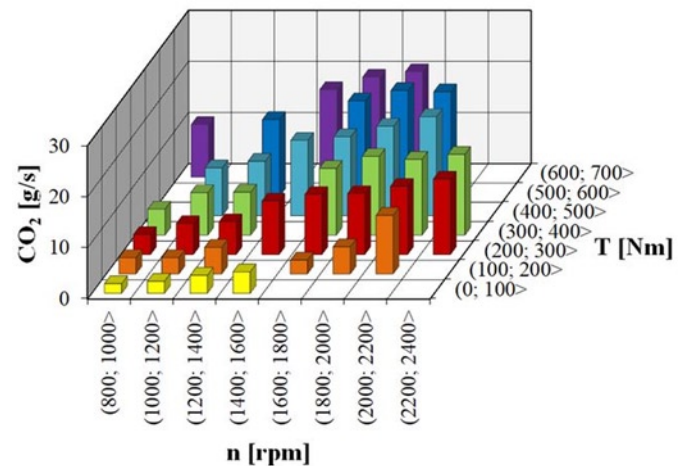


Fig. 6. The CO_2 exhaust emission intensity characteristics in the engine speed-load intervals

For the presented CO emission intensity characteristic (Fig. 7a), the highest values occurred in the interval (1400; 2000 rpm) and (500; 700 Nm), with the local maximum being 0.09 g/s. The average value for this interval was 0.06 g/s. One should also note the rotational speeds interval of (1600; 2400 rpm), in which the mean emission intensity value was 0.05 g/s. The obtained values resulted from the operating conditions during which the fuel dose was increased, which causes global and local oxygen deficiencies closely related to the formation of carbon oxides, their formation was also influenced by the high temperature in the combustion chamber (locally exceeding 2000°C in farm tractors), which was the conditions favouring the thermal dissociation into CO. This reaction occurs at temperatures greater than 2000 K, while above 3000 K 40% of carbon dioxide dissociates into CO.

The mean exhaust emission value for the entire test drive was 0.004 g/s. The values of the $M_{\text{CO}}/\text{CO}_2$ emission indicator were evenly distributed throughout the entire engine operating range and remained in the range between 1.7–6.1 (Fig. 7b). The maximum value of 6.1 was obtained for a single operating point in the intervals of rotational speed (1400; 1600 rpm) and load (200; 300 Nm). The mean emission indicator value was 3.6.

Characterization of the hydrocarbon emission intensity as a function of engine speed and load (Fig. 8a) showed that the highest value was obtained in the area within the rotational speed interval (1600; 2200 rpm) and load interval (400; 700 Nm), for which the average remained at the level of 0.004 g/s. The maximum local value (0.005 g/s) was recorded in a single measurement interval for (1600; 1800 rpm) and (600; 700 Nm). The mean value of hydrocarbon exhaust emission intensity over the entire test drive was 0.003 g/s. The increase in HC emission rate was caused by high rotational speed, at which the injected dose of fuel is not thoroughly mixed with the air, which leads to incomplete combustion. Most of the factors contributing to the formation of an excessive amount of carbon monoxide in the exhaust gas also cause excessive emission of hydrocarbons, hence the local maxima of both compounds were recorded in the same intervals of the engine operating points. The highest values of $M_{\text{HC}}/\text{CO}_2$ emission indicators (0.73) were obtained for the rotational speed range (800; 1000 rpm) at a load of (0; 100 Nm), i.e. for engine idling, where the exhaust aftertreatment systems had not achieved their light-off temperature, and when the overall combustion temperature was low. The

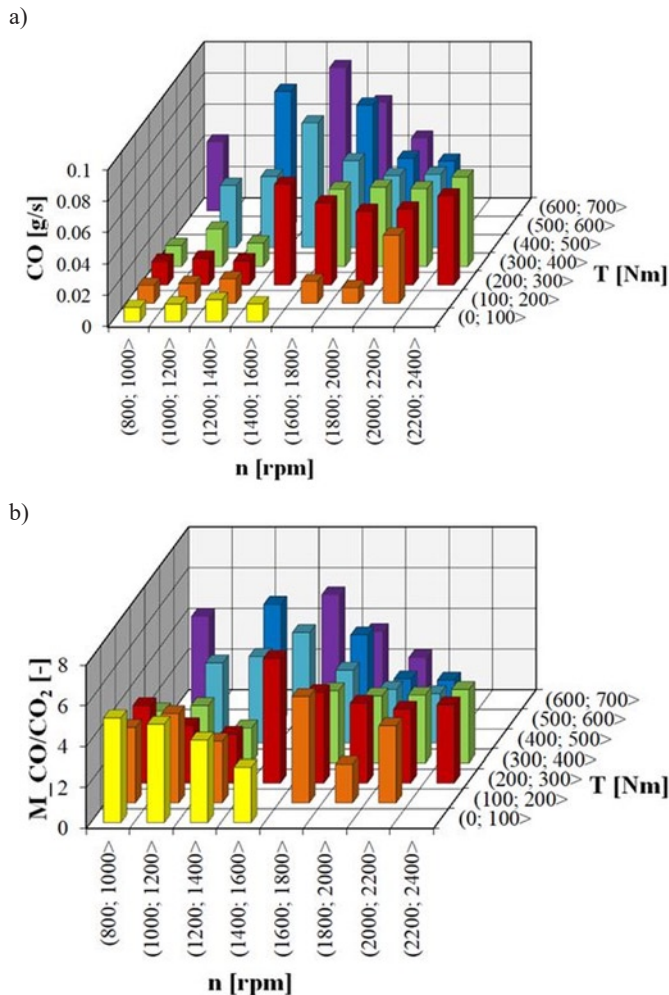


Fig. 7. Characteristics of a) CO emission per second; b) CO emissivity in the torque ranges of the engine speed

mean value of the emission values for the entire test was 0.035 g/s. The values of the discussed exhaust compound were evenly distributed throughout the engine operation area, similarly to CO emission indicators (Fig. 8b).

The highest NO_x emission intensity was recorded for the engine rotational speed interval (1600; 2200 rpm) and load interval (400; 700 Nm) - Fig. 9a, for which the mean value was 0.053 g/s. The maximum value of 0.1 g/s occurred in two separate measurement intervals (1800; 2000 rpm) and (500; 600 Nm) as well as for (2000; 2200 rpm) and (500; 600 Nm).

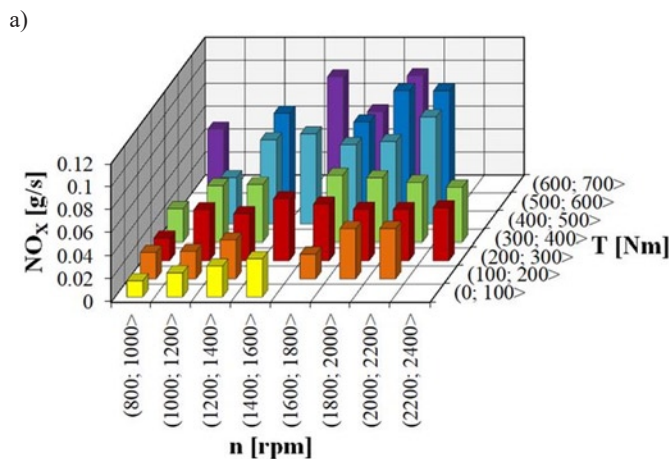


Fig. 9. Characteristics of a) second NO_x emission rate b) NO_x toxicity in the torque ranges of the engine speed

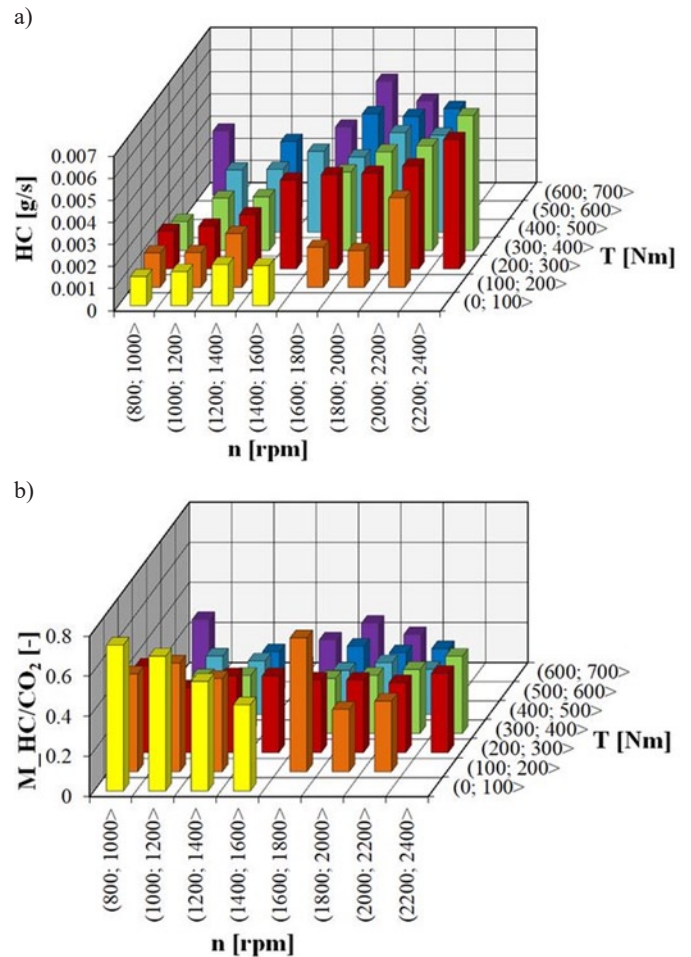
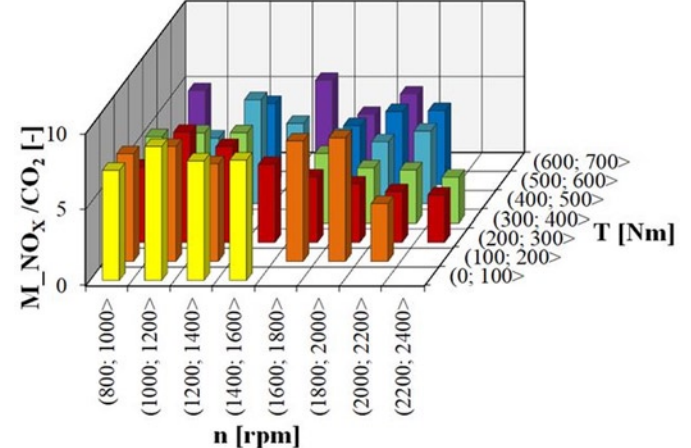


Fig. 8. Characteristics of a) second HC emission rate; b) HC emissivity in the torque ranges of the engine speed

At significant engine speeds, above 800 rpm, the engine generates a higher cylinder temperature, which directly promotes the formation of nitrogen oxides. Distribution of the exhaust emission characteristics is similar to the distribution of the previously described exhaust compounds, and therefore uniform in the entire range of the engine operation. The emission indicator values varied in the range of 3–8.8, with the local maximum 8.8 being recorded for engine parameters enclosed in the intervals of rotational speed (100; 1200 rpm) and load (0; 100 Nm). The mean value for the entire operating area of parameter variability was 5.4 (Fig. 9b).

Based on the determined masses of individual emitted gaseous compounds, a comparison of their emission indicators was made for



each test sample, which differed in the vehicle speed value (5 km/h, 10 km/h, 15 km/h). For each exhaust compound, the lowest emission rates were found for the test characterized by the highest speed of 15 km/h. These values were 2.42, 0.31, 4.86 for CO, HC and NO_x respectively. In the first test, there was an almost 70% difference in the HC emission indicator value (compared to the third trial) and a 30% difference in the indicator for NO_x. The second test (at 10 km/h) was characterized by the highest emission indicator for CO, and the differences in the values were 34% and 32% respectively when compared to the tests performed at 5 km/h and 15 km/h (Fig. 10). The measurement results presented in [15] showed that increasing the vehicle speed when performing field work may contribute to the reduction of fuel consumption. During the tests described in [15], an 18% reduction in fuel consumption was achieved as a result of increasing the drive speed with the field cultivator from 5 km/h to 15 km/h. Fuel consumption tests were carried out in the same field in successive cycles, i.e. the measurements were made for the same type of soil and for the same weather conditions. Fuel consumption is closely related to the emission of harmful CO₂ and other toxic exhaust gas compounds, hence it confirms that the lowest emission indicators were recorded in the third measurement test.

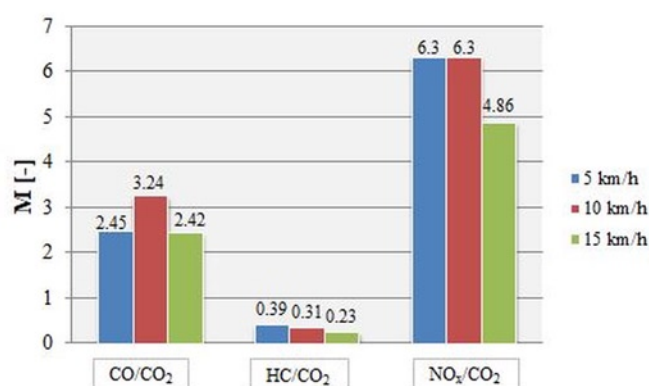


Fig. 10. Comparison of M emission indicators for CO, HC and NO_x obtained during the three test drives

6. Conclusion

General conclusion:

- The paper proposes a new proprietary emission factor M, based on the use of measurements of carbon dioxide parameters (e.g. the emission intensity of this compound, its road or unit emission). The values that are substituted into its structure must be expressed in the same units, which will make it dimensionless.
- The conducted analyzes, in particular the introduced emission indicator, are a novelty in the aspect of emissivity analysis.

Detailed conclusions:

- On this basis of research, it was found that the lowest values of the emission indicators M were recorded for the test characterized by the highest vehicle speed of 15 km/h. Therefore, increasing the driving speed of the tractor during typical field work from 5 km/h to 15 km/h may have a positive impact on the overall exhaust emission results of toxic compounds.
- In the previous work of the authors [17], considerations were made in terms of the possible applications of the emission indicator for conventional, hybrid and alternative fuel urban buses.

References

1. Almén J. Swedish In-Service Testing Programme on Emissions from Heavy-Duty Vehicles. AVL Sweden Certification & Regulation Compliance. 2010.
2. Anser MK, Apergis N, Syed QR. Impact of economic policy uncertainty on CO₂ emissions: evidence from top ten carbon emitter countries. Environmental Science and Pollution Research 2021; 1-10, <https://doi.org/10.1007/s11356-021-12782-4>.
3. Association for Emissions Control by Catalyst (AECC). Workshop on Clean Air and Real Driving Emissions. Conference materials of: Motor Transport Institute in Warsaw 2010.

Measurements were made in accordance with the SORT 2 drive test procedure and on the test route in the Poznań agglomeration. The comparison of the emission indicators from both studies shows that the NRMM vehicle at a speed of 10 km/h is characterized by similar trends in the values of the M indicators as the hybrid bus. In the case of CO/CO₂, the vehicle achieved exactly the same value (3.24) as the hybrid bus during road tests on the city line and 3.44 for a trip in the standardized SORT 2 test. This is due to the similar performance characteristics of the engines of both these vehicles, namely the engine operating at higher loads (above 50%). The comparison of the values of the remaining emission indicators for the NRMM vehicle (during the 10 km/h test) with the hybrid bus results in the following values:

- $M_{HC_NRMM} = 0.31$; $M_{HC_urban\ route} = 0.24$; $M_{HC_SORT\ 2} = 0.2$,
- $M_{NOx_NRMM} = 6.3$; $M_{NOx_urban\ route} = 9.63$; $M_{NOx_SORT\ 2} = 6.97$.

The comparative analysis of the agricultural tractor and the hybrid bus also showed that in both cases the highest values of the M_{NOx} indicator were achieved, which, as already mentioned, results from the operation of the engines in the higher efficiency range of operating points. It should be noted that the drives of the test vehicles had different rated parameters, and the hybrid vehicle was equipped with an SCR aftertreatment system. Similar considerations were also carried out in [16], where a comparative analysis of the emission indicator for two passenger cars and a city bus was performed. Therefore, this analysis confirms the universality of the M indicator as a tool for comparison of overall engine.

Methodological conclusions:

- It is necessary to determine the resistances of the internal combustion engine and take into account their values when determining the torque in tests carried out during the vehicle operation and the work performed by the drive system.
- Many studies have shown [13] that the current type approval tests (static - NRSC and dynamic - NRTC) do not fully reflect both the actual engine operation parameters and the emission of toxic exhaust compounds. Therefore, it is necessary to continue work on testing the exhaust emission of toxic compounds from this group of vehicles and changing the regulations regarding emission control strategies. In addition, an important aspect in relation to the vehicles in the NRMM group is also work on the improvement of mobile exhaust emission measuring equipment in order to optimize the research process itself with this type of machines.

Prognostic conclusions:

- The literature review allows the authors of this work to state that its subject matter is consistent with the direction of research carried out around the world, and the presented emission index M met with the approval of the scientific world and representatives of the European Commission at numerous industry conferences.
- In the longer term, taking into account the dimensionlessness of the M emission indicator, it is possible to compare not only vehicles of the same category, but also objects belonging to different groups, from LDV to NRMM vehicles. Thus, the dissemination of the indicator presented allows for even more precise definition of the environmental performance of a vehicle / machine.

4. Barlow TJ, Latham S, McCrae IS, Boulter PG. A reference book of driving cycles for use in the measurement of road vehicle emissions. TRL Published Project Report. 2009.
5. Čupera J, Sedlák P. Design and verification of engine power calculation model using the data of a digital bus built into an agricultural tractor. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis* 2014; 59(6): 111-120, <https://doi.org/10.11118/actaun201159060111>.
6. Daszkiewicz P, Andrzejewski M. Preliminary analyzes in terms of the possibility of reducing energy consumption by the SM42 locomotive used in track works. *MATEC Web of Conferences*. 2017; 118(00014), <https://doi.org/10.1051/mateconf/201711800014>.
7. Fuc P, Lijewski P, Kurczewski P, Ziolkowski A, Dobrzynski M. The analysis of fuel consumption and exhaust emissions from forklifts fueled by diesel fuel and liquefied petroleum gas (LPG) obtained under real driving conditions. *ASME International Mechanical Engineering Congress and Exposition 2017. American Society of Mechanical Engineers Digital Collection*, <https://doi.org/10.1115/IMECE2017-70158>.
8. Fuc P, Lijewski P, Ziolkowski A, Dobrzynski M. Development of a method of calculation of energy balance in exhaust systems in terms of energy recovery. *ASME International Mechanical Engineering Congress and Exposition 2017*; 58431(V008T10A047), <https://doi.org/10.1115/IMECE2017-70159>.
9. Fuc P, Lijewski P, Ziolkowski A. Analysis of the CO₂, NO_x emission and fuel consumption from a heavy-duty vehicle designed for carriage of timber. *IOP Conference Series: Materials Science and Engineering* 2016; 148 (1): 012065, <https://doi.org/10.1088/1757-899X/148/1/012065>.
10. Giechaskiel B, Lähde T, Gandi S, Keller S, Kreutziger P, Mamakos A. Assessment of 10-nm particle number (PN) portable emissions measurement systems (PEMS) for future regulations. *International Journal of Environmental Research and Public Health* 2020; 17(11): 3878, <https://doi.org/10.3390/ijerph17113878>.
11. International Council on Clean Transportation. European Stage V non-road emission standards. 2016;1-8.
12. Kuranc A. Exhaust emission test performance with the use of the signal from air flow meter, *Eksplotacja i Niezawodność - Maintenance and Reliability* 2015; 17 (1): 129-134, <https://doi.org/10.17531/ein.2015.1.17>.
13. Lijewski P. Study of exhaust emission from non-road engines. Dissertation, Poznan University of Technology 2013.
14. Lijewski P, Fuc P, Dobrzynski M, Markiewicz F. Exhaust emissions from small engines in handheld devices. *MATEC Web of Conferences* 2017; 118(00016), <https://doi.org/10.1051/mateconf/201711800016>.
15. Merksiz J, Lijewski P, Fuc P, Siedlecki M, Weymann S. The use of the PEMS equipment for the assessment of farm fieldwork energy consumption. *Applied Engineering in Agriculture* 2015; 31(6), <https://doi.org/10.13031/aea.31.11225>.
16. Merksiz J, Rymaniak Ł. Determining the environmental indicators for vehicles of different categories in relation to CO₂ emission based on road tests. *Combustion Engines* 2017; 56, <https://doi.org/10.19206/CE-2017-310>.
17. Merksiz J, Rymaniak Ł. The assessment of vehicle exhaust emissions referred to CO₂ based on the investigations of city buses under actual conditions of operation. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2017; 19(4): 522-529, <https://doi.org/10.17531/ein.2017.4.5>.
18. Merksiz J, Waligorski M, Bajerlein M, Markowski J. Application of the Frequency and JTFA Analyses of the Accompanying Processes for OBD Combustion Process Monitor Design in Turbocharged CI Direct Injection Engines of HDV Non-Road Vehicles. *SAE Technical Paper* 2011, <https://doi.org/10.17531/ein.2017.4.5>.
19. Regulation (EU) 2019/631 of the European Parliament and of the Council of 17 April 2019 setting CO₂ emission performance standards for new passenger cars and for new light commercial vehicles and repealing Regulations (EC) No 443/2009 and (EU) No 510/2011.
20. Regulation (EU) 2016/1628 of the European Parliament and of the Council of 14 September 2016 on requirements relating to gaseous and particulate pollutant emission limits and type-approval for internal combustion engines for non-road mobile machinery, amending Regulations (EU) No 1024/2012 and (EU) No 167/2013.
21. Siedlecki M, Lijewski P, Weymann S. Analysis of tractor particulate emissions in a modified NRSC test after implementing a particulate filter in the exhaust system. *MATEC Web of Conferences* 2017; 118(00028), <https://doi.org/10.1051/mateconf/201711800028>.
22. Warguła Ł, Kukla M, Lijewski P, Dobrzyński M, Markiewicz F. Influence of Innovative Woodchipper Speed Control Systems on Exhaust Gas Emissions and Fuel Consumption in Urban Areas. *Energies* 2020; 13(13): 3330, <https://doi.org/10.3390/en13133330>.
23. www.semtech.com

Tool wear condition monitoring in milling process based on data fusion enhanced long short-term memory network under different cutting conditions

Indexed by:



Guoxiao Zheng^a, Weifang Sun^a, Hao Zhang^b, Yuqing Zhou^{a,*}, Chen Gao^{c,*}

^aCollege of Mechanical and Electrical Engineering, Wenzhou University, Wenzhou, China, 325035

^bShaoxing Customs, Shaoxing, China, 312099

^cSchool of Mechatronics and Transportation, Jiaxing Nanyang Polytechnic Institute, Jiaxing, China, 314031


Highlights

- A data fusion- LSTM is proposed to estimate tool wear under different cutting conditions.
- NCA is used to select useful features fused by EMD VMD and FSST.
- Experimental results show the proposed method outperforms significantly SVR and RNN.

Abstract

Tool wear condition monitoring (TCM) is essential for milling process to ensure the machining quality, and the long short-term memory network (LSTM) is a good choice for predicting tool wear value. However, the robustness of LSTM- based method is poor when cutting condition changes. A novel method based on data fusion enhanced LSTM is proposed to estimate tool wear value under different cutting conditions. Firstly, vibration time series signal collected from milling process are transformed to feature space through empirical mode decomposition, variational mode decomposition and fourier synchro squeezed transform. And then few feature series are selected by neighborhood component analysis to reduce dimension of the signal features. Finally, these selected feature series are input to train the bidirectional LSTM network and estimate tool wear value. Applications of the proposed method to milling TCM experiments demonstrate it outperforms significantly SVR- based and RNN- based methods under different cutting conditions.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

tool wear condition monitoring, empirical mode decomposition, variational mode decomposition, fourier synchro squeezed transform, neighborhood component analysis, long short-term memory network.

1. Introduction

In the modern numerical control milling process, tool condition is one of the key factors affecting the machining quality of workpiece [19, 22]. Tool breakage is the main cause of abnormal shutdown and lead to time lost and capital destroyed [27]. It has reported that severe tool failure causes at least 20% of abnormal downtime [4, 32]. However, traditional tool condition monitoring (TCM) methods are based on the machining time or the number of workpiece machined resulting in the effective utilization rate of tool is only 50%-80%, which affect the processing efficiency and increase the machining cost significantly [15, 35]. It is predicted that an effective TCM method can increase the cutting efficiency by 10-50% and reduce the machining cost by 10-40% [23, 33]. Therefore, the development of effective online TCM method has received broadly positive reviews and is a research hotspot nowadays [10, 11].

Recently, many deep learning models have been employed in TCM applications [9, 14, 21]. For example, Cao et al [1] recognized tool wear condition by derived wavelet frames and Convolutional

neural network (CNN) using vibration signals. Recent advanced technology that have greatly increased the number of TCM study, Huang et al [8] proposed a tool wear predicting method by deep CNN, in which multi- domain features are respectively extracted from cutting force and vibration. Lei et al [16] employed Extreme learning machine (ELM) to classify tool wear condition in milling processes, and used genetic algorithm and particle swarm optimization to optimize model parameters of ELM. Tim and Chris [26] proposed a disentangled- variation- autoencoder CNN method to estimate tool wear condition in a self-supervised way. Zhi et al [30] proposed a hybrid CNN and edge-labeling graph neural network (EGNN) method for limited tool wear image training samples, in which the CNN is employed to extract features of tool wear image and the EGNN is applied to distinguish the tool's category. However, these TCM methods have been generally applied for diagnosis (classification) rather than prognosis (regression), tool wear is a progressive and continuous cumulative process, regressive prediction of tool wear is more suitable than classification that the CNN- based methods are difficult to use [34]. Recurrent neural networks (RNN) could be solve the problem

(*) Corresponding author.

E-mail addresses: G. Zheng - 909936695@qq.com, W. Sun - vincent_suen@126.com, H. Zhang - jiyongxing@customs.gov.cn, Y. Zhou - zhuyq@wzeducn, C. Gao - gaochen_1993@163com

of regression and increase the accuracy of the prognosis, but the error of backpropagation in RNN would increase sharply or decrease exponentially, which lead to the problem of long lag [5,18]. As a significant branch of RNN, Long short-term memory (LSTM) network is proposed to overcome the above problem. Due to the special unit structure with learning long-term dependencies, LSTM can deal with the long-distance dependence problem in time sequence data [6]. Therefore, LSTM is potential to obtain good performance for TCM [31]. Tao et al [24] designed a TCM method based on LSTM and hidden Markov model (HMM) to estimate the tool wear value and predict its remaining useful life. Zhao et al [29] proposed a convolutional Bi-directional LSTM network, in which CNN extracted local feature of original signal and Bi-directional LSTM encoded temporal information and predict tool wear value. However, it is found that the regression accuracies of LSTM- based TCM method are poor when the cutting conditions of testing samples are different with that of training samples in our experiment. That is, the cutting condition could affect significantly the performance of LSTM- based TCM method. Therefore, this paper try to alleviated the influence of cutting condition to LSTM model through a data fusion way.

In this paper, a data fusion enhanced LSTM- based TCM method is established to estimate tool wear value under variable cutting conditions. The paper is organized as follows: Section 2 introduces the proposed data fusion enhanced LSTM method, Section 3 describes the experimental setup, data analysis and experimental results. Finally, conclusion is in Section 4.

2. Proposed method

2.1. Framework of the proposed method

The proposed TCM method framework based on data fusion enhanced LSTM is illustrated in Figure 1. Firstly, vibration time series signal collected from milling process are transformed to feature space through Empirical mode decomposition (EMD), Variational mode decomposition (VMD) and Fourier synchro squeezed transform (FSST), and then few feature series are selected by neighborhood component analysis (NCA) to reduce dimension of the signal features. Finally, these new feature series selected by NCA are input into bidirectional LSTM network to train the regression model.

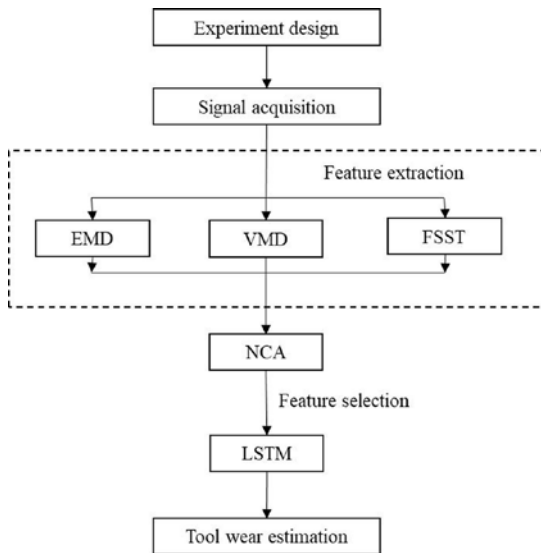


Fig. 1. Framework of the data fusion enhanced LSTM- based TCM method

2.3. Data preprocessing

For extracting more features of time series under limited samples, the collected signals are divided into multiple segments using a sliding window method. In addition, these segmented data are normalized by batch normalization method [17] as follows:

$$\mu_B = \frac{1}{m} \sum_{i=1}^m x_i \quad (1)$$

$$\sigma_B^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2 \quad (2)$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \quad (3)$$

where x_i and y_i denote the input and output value after batch normalization respectively, m denotes the number of inputs in minibatch, μ_B and σ_B denote the mean of input and the average variance of the input respectively, \hat{x}_i is the normalized x_i .

2.3. Feature extraction

2.3.1. Empirical mode decomposition

EMD is a nonlinear time-frequency decomposition algorithm that decompose the signal into several intrinsic mode functions (IMFs) and a residual [7], shown in Equation (4). In EMD, all decomposed IMFs contain the local feature information in different time scales of the original signal. Finally, each IMF contains approximately a single frequency component, and the instantaneous frequency of the original signal can be obtained after the weighted average of the instantaneous frequency of each IMF:

$$X(t) = \sum_{i=1}^N \text{IMF}_i(t) + r_N(t) \quad (4)$$

EMD decomposes the signal according to the time scale features of the original data, without pre-setting any basis function, which is the most significant advantage compared with other time-frequency decomposition methods, such as wavelet transform. Due to the complexity and uncertainty of milling process, it is very difficult to find a basis function suitable for milling signal, EMD could be employed for feature extraction in milling TCM.

2.3.2. Variational mode decomposition

VMD is an adaptive time-frequency signal decomposition algorithm, its framework is the solution of variational problems [3]. VMD considers the signal is composed of sub signals with different frequencies dominant, and transforms the decomposition of signal into the solution of constrained variational model [13,28]. In this process, the central frequency and bandwidth of each IMF are updated alternately and iteratively. Finally, the signal band is decomposed adaptively and obtain the preset K narrowband IMFs in equation (5).

$$x(t) = \sum_{k=1}^K u_k(t) \quad (5)$$

In VMD, each IMF u_k is a bandwidth limited frequency modulation and amplitude modulation signal shown in equation (6):

$$u_k(t) = A_k(t) \cos(\phi_k(t)) \quad (6)$$

VMD has perfect mathematical theory support, its essence is an adaptive optimal Wiener filter group, which can get high signal-to-noise ratio IMFs.

2.3.3. Fourier synchro squeezed transform

Fourier synchro squeezed transform (FSST) is based on the short-time Fourier transform (SFT) implemented in the spectrogram function [12,25]. The FSST function determines the SFT of a function, f using a spectral window, g , and computing in equation (7):

$$V_g f(t, \eta) = \int_{-\infty}^{\infty} f(x) g(x-t) e^{-j2\pi\eta(x-t)} dx \quad (7)$$

Unlike the conventional definition, this definition has an extra factor of $e^{j2\pi\eta t}$. The transform values are then “squeezed” so that they concentrate around curves of instantaneous frequency in the time-frequency plane.

2.3.4. Neighborhood component analysis

Neighborhood component analysis (NCA) is a distance metric method in metric learning and dimension reduction fields [2]. NCA is based on K-Nearest Neighborhood (KNN) including feature parameters and response label [20]. NCA selects randomly neighbors, obtains the transformation matrix in Mahalanobis distance by optimizing the results of the leave-one-out cross validation (LOOCV) method, and finds the feature parameter set maximizing the average LOO classification / regression accuracy to achieve the purpose of feature selection.

2.4. Long short-term memory network

An LSTM network is a type of RNN that can learn long-term dependencies between time steps of sequence data [6,29]. The framework of LSTM is shown in Figure 2.

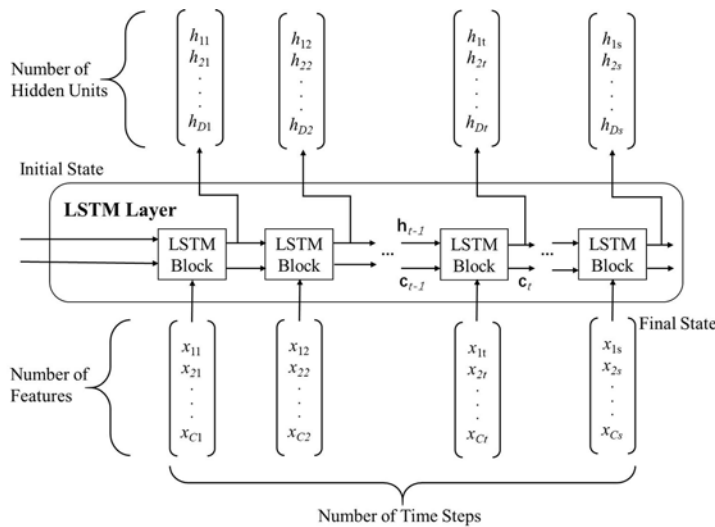


Fig. 2. LSTM architectures

Table 1. Definition and expression of the LSTM layer gate

| Component | Purpose | Formula |
|------------------------|---|---|
| Input gate (i) | Control level of cell state update | $i_t = \sigma_g(W_i \mathbf{x}_t + R_i \mathbf{h}_{t-1} + b_i)$ |
| Forget gate (f) | Control level of cell state reset (forget) | $f_t = \sigma_g(W_f \mathbf{x}_t + R_f \mathbf{h}_{t-1} + b_f)$ |
| Cell candidate (g) | Add information to cell state | $g_t = \sigma_c(W_g \mathbf{x}_t + R_g \mathbf{h}_{t-1} + b_g)$ |
| Output gate (o) | Control level of cell state added to hidden state | $o_t = \sigma_g(W_o \mathbf{x}_t + R_o \mathbf{h}_{t-1} + b_o)$ |

where W_t and R_t are the input weights and recurrent weight in the t -th layer, and b_k is the bias of each component.

Let $X_t = \{X_{1t}, X_{2t}, \dots, X_{Ct}\}$ is a time series with C features, \mathbf{h}_t and \mathbf{c}_t are the hidden state and cell state at time t , respectively. At time t , the state of the network ($\mathbf{c}_t, \mathbf{h}_t$) is calculated by X_t and $(\mathbf{c}_{t-1}, \mathbf{h}_{t-1})$ by Equation (8) and (9):

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot g_t \quad (8)$$

$$\mathbf{h}_t = o_t \odot \sigma_c(\mathbf{c}_t) \quad (9)$$

The definition and expression of it fit gt ot are as shown in Table 1.

3. Experimental observation and research

3.1. Experimental setup

The experimental device for milling TCM is shown in Figure 3. In the milling TCM experiment, a CNC milling machine (DMTG VDL850A, China) is used to finish milling process, and a piece of #45 steel (30 cm × 10 cm × 8 cm) is used as the workpiece material. What's more, the milling vibration signals of spindle X and Y directions are acquired by two accelerometers with a signal acquisition device (ECON Dynamic Signal Analyzer, shown in Figure 3(b)). In addition, the signal sampling frequency in the experiment is 12KHz.

Fourteen uncoated three-insert tungsten steel end milling cutters with diameter of 10 mm are employed to mill the workpiece under different cutting conditions, listed in Table 2. For each tool, the workpiece is milled surface 10 times, and the tool wear value is measured after milling each surface using a tool microscope (GP-300C Figure 3(c)). The length of rake face wear (KB) is employed as the tool wear criterion in the experiment, and the max value $KB = \max(KB_1, KB_2, KB_3)$ of three inserts is adopted as the final tool wear value. Figure 4 illustrates the tool wear conditions after milling the workpiece surface 1-st, 5-th and 10-th times.

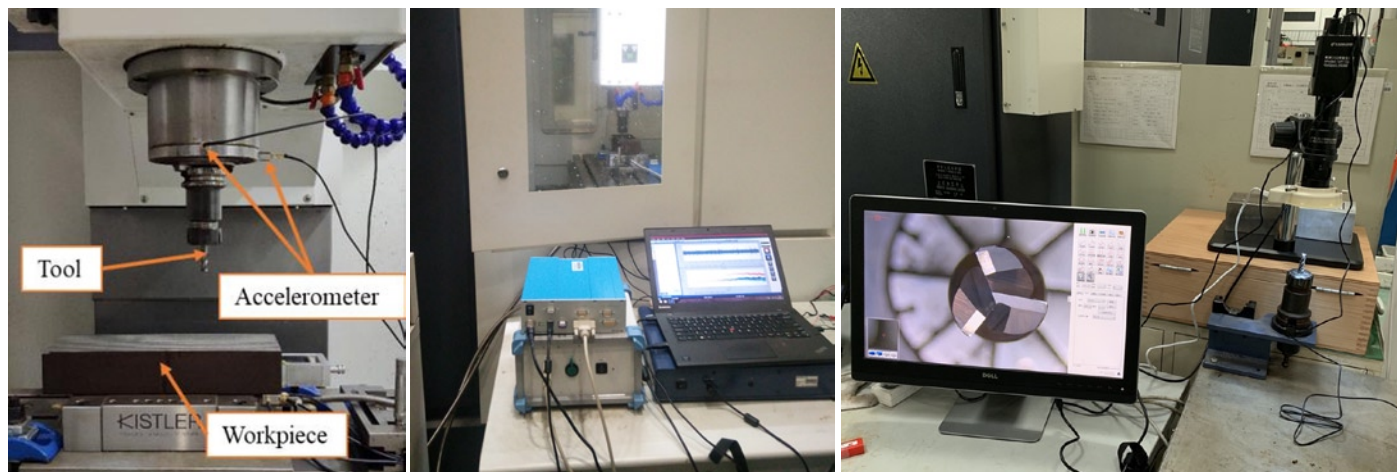
In the 14 milling TCM experiments, the training, verification and testing sets are generated randomly shown in Table 2, 7 sets of samples for training, 3 sets of samples for verification, and 4 sets of samples for testing.

3.2. Results and analysis

3.2.1. Samples and metrics

Acceleration signals of Spindle X and Y direction are used in the network, 272 training set, 120 validations set, and 80 test set are made up from spindle sensor signals. In all analyzed samples, there is no same cutting condition combination in the three dataset. Besides, in the signal pre-processing, the original signal of each sample is divided into 10 parts by slide window method, in which the window size is 2000 points, and the sliding distance is 1000 points.

To evaluate the performance of the proposed method, three indexes are employed, including the mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2).

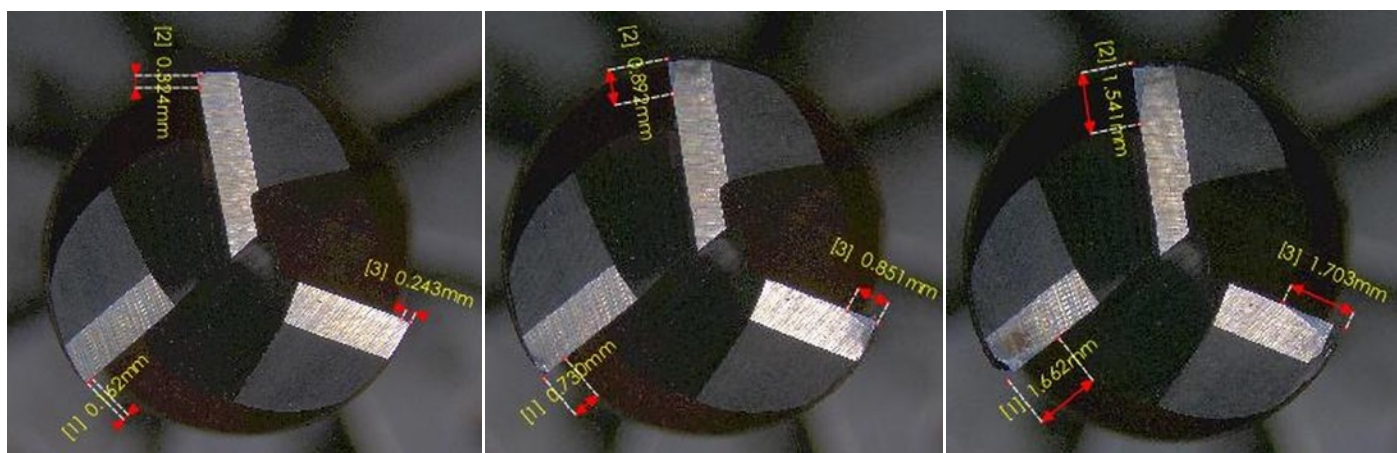


(a) Experimental platform

(b) Data acquisition system

(c) Tool microscope

Fig. 3. The experimental setup [16]



(a) 1-st milling

(b) 5-th milling

(c) 10-th milling

Fig. 4. Tool images indicative of different tool-wear values [33]

Table 2. Experimental cutting parameters

| Case No | Spindle speed (rpm) | Axial cut deep (mm) | Feed speed (mm/min) | Dataset type |
|---------|---------------------|---------------------|---------------------|--------------|
| 1 | 2300 | 0.4 | 400 | Training |
| 2 | 2300 | 0.5 | 450 | Validation |
| 3 | 2300 | 0.6 | 500 | Testing |
| 4 | 2400 | 0.4 | 450 | Training |
| 5 | 2400 | 0.5 | 500 | Testing |
| 6 | 2400 | 0.6 | 400 | Validation |
| 7 | 2500 | 0.4 | 500 | Training |
| 8 | 2500 | 0.5 | 400 | Testing |
| 9 | 2500 | 0.6 | 450 | Training |
| 10 | 2300 | 0.4 | 500 | Testing |
| 11 | 2300 | 0.6 | 400 | Training |
| 12 | 2500 | 0.6 | 500 | Validation |
| 13 | 2500 | 0.6 | 400 | Training |
| 14 | 2500 | 0.4 | 400 | Training |

3.2.2. Algorithm settings

For each cutting process in the experiment, there are two mutually perpendicular milling vibration signals which are collected from the equipment and a part of collected signal has 12000 points as shown in Figure 5, in which the corresponding cutting parameters is the Case 1

in Table 2: spindle speed is 2300 rpm, axial cutting depth is 4 mm, and feed rate is 400 mm/min.

Since the real monitoring signal is often nonlinear and non-stationary, it is suitable to use the EMD, VMD and FSST methods to obtain the features of vibration signals for tool wear. In order to obtain signal

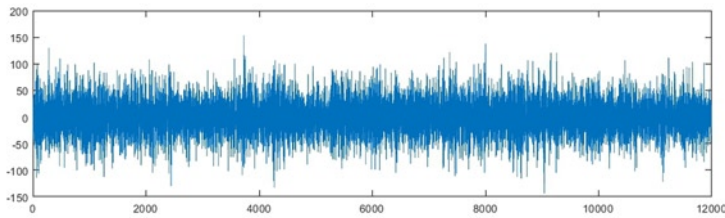


Fig. 5. Original vibration signal

Table 3. Features selection

| Characteristic | Fourier synchro squeezed transform(Hz) |
|---|--|
| Real part characteristic frequency | 0;938;1875;2813;375;4688;5625;6563;750;8438;9375;103130;1125;12188;1312514063;4500(Hz) |
| Imaginary part characteristic frequency | 938;1875;2813;375;4688;5625;6563;750;8438;9375;10313;1125;12188;4500(Hz) |

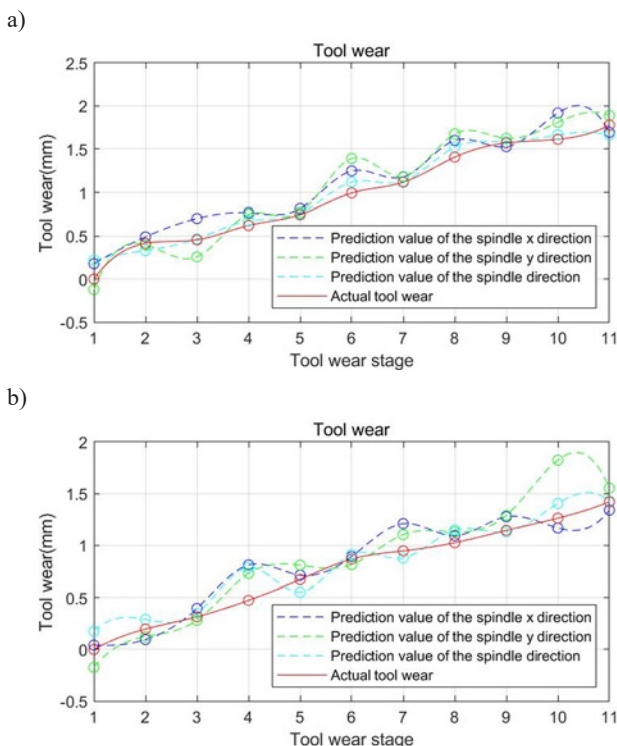


Fig. 6. Prediction results of tool wear: a) the 5-th tool, b) the 8-th tool

feature and more information from the vibration signal to predict the tool wear value, the original signal is transformed by EMD, VMD and FSST to expand the dimensionality. Furthermore, to remove irrelevant features and reduce the number of features, sensitive features that correlate well with tool wear are selected out through NCA.

Table 4. Network architectures

| Serial number | Name | Type | Serial number | Name | Type |
|---------------|---------------|--------------------------|---------------|-------------------|------------------------------|
| 1 | sequenceinput | Sequence input | 7 | dropout_2 | 20% dropout |
| 2 | biLSTM_1 | BILSTM: 300 hidden units | 8 | fc_1 | 1 fully connected layer |
| 3 | relu_1 | ReLU | 9 | dropout_3 | 20% dropout |
| 4 | dropout_1 | 20% dropout | 10 | fc_2 | 1 fully connected layer |
| 5 | biLSTM_2 | BILSTM: 300 hidden units | 11 | Regression output | mean-squared-error: Response |
| 6 | relu_2 | ReLU | | | |

In this work, the first 6 IMFs and residuals are taken in EMD, the first 5 IMFs and residuals are taken in VMD, and 60 IMFs are decomposed in FSST. In addition, it is necessary to take the real and imaginary parts of the IMF as the feature matrix of the vibration signal, and use NCA to take the effective characteristic matrix. The results are listed in Table 3

By calculating, it was found that the feature matrix has first 6 numbers of IMFs and residuals of EMD, 5 numbers of IMFs and residuals of VMD, 17 real parts and 14 imaginary parts of FSST. Totally 45 feature matrixes as input signals. The two single-channel experimental data of the sensor are superimposed and fused into a new sample. Meanwhile, all data from experiments need batch normalization.

In this model, it is a way to use eleven layers as neural network architectures in our experiments: especially bidirectional LSTM layer, which has two hidden LSTM layers (forwards and backwards) as shown in Table 4.

Due to the limitations of experimental equipment conditions and cost, 14 sets of experiments were executed, 7 sets of samples under different working conditions were selected for training, 3 sets of samples were selected for verification, and 4 sets of samples were selected for testing.

For all architectures, complete error gradient was calculated and the weights are trained by using gradient descent with momentum. In all experiments, the same training parameters were kept: randomly assigned initial weights, keeping the training algorithm and parameters constant, allowing us to focus on the impact of changing the architecture.

3.2.3. Experimental results

The LSTM model established by the training set and verification set is applied to predict the testing set, including 4 tools with different cutting conditions. In Figure 6, the blue, green, and cerulean dotted lines denote the prediction results using the proposed method with the spindle vibration signal of X-direction, Y-direction, and dual-direction (composition of X and Y directions). It is noted that the cutting parameters of the 5-th and 8-th tools are different. For Figure 6(a), the spindle speed is 2400 rpm, the axial cutting depth is 0.6 mm, and the feed rate is 500 mm/min. For Figure 6(b), the spindle speed is 2500 rpm, the axial cutting depth is 0.5 mm, and the feed rate is 400 mm/min. It can be seen that the trend of the overall predicted value is similar to the actual wear value, and the error at some stages is less than 0.1 or even close to the wear value.

To test the regression performance, the proposed method is compared with RNN and support vector machine (SVR). As a result, the MSE, RMSE and R^2 of three methods are presented in Table 5.

It can be seen from Table 5 that the proposed LSTM-based method is highly effective in improving the regression accuracy, the prediction accuracy of the proposed method is much higher than that of RNN and SVR according to the values of three evaluation indexes, except for the X-direction signal of the 3-rd and 5-th tools. In addition, the prediction accuracies with the dual-direction signal outperform that of signal-direction except for the 3-rd tool, while the results of three indexes are slightly worse than that of two other methods in the 3-rd tool.

Table 5. Prediction results of LSTM and RNN and SVR

| Vibration signal | Tool | MAE | | | RMSE | | | R-Squared | | |
|-------------------------|------|---------------|--------|---------------|---------------|--------|---------------|---------------|---------|---------------|
| | | LSTM | SVR | RNN | LSTM | SVR | RNN | LSTM | SVR | RNN |
| Spindle X- direction | 3 | 0.2150 | 0.5091 | 0.1222 | 0.2572 | 0.6160 | 0.1490 | 0.7996 | -0.1498 | 0.9328 |
| | 5 | 0.1508 | 0.5518 | 0.1127 | 0.1741 | 0.6443 | 0.1396 | 0.9091 | -0.2456 | 0.9415 |
| | 8 | 0.1153 | 0.4822 | 0.2507 | 0.1483 | 0.5824 | 0.2669 | 0.8971 | -0.5871 | 0.6667 |
| | 10 | 0.1762 | 0.4686 | 0.3688 | 0.2063 | 0.6246 | 0.4857 | 0.8039 | -0.7978 | -0.0868 |
| Spindle Y- direction | 3 | 0.2829 | 0.3989 | 0.2572 | 0.3327 | 0.4891 | 0.4811 | 0.6647 | 0.2829 | 0.3989 |
| | 5 | 0.1437 | 0.6037 | 0.2458 | 0.1813 | 0.7576 | 0.4740 | 0.9014 | 0.1437 | 0.6037 |
| | 8 | 0.1661 | 0.5130 | 0.6122 | 0.2159 | 0.7089 | 0.7402 | 0.7819 | 0.1661 | 0.5130 |
| | 10 | 0.1681 | 0.5639 | 0.6996 | 0.1985 | 0.6385 | 0.7295 | 0.8184 | 0.1681 | 0.5639 |
| Spindle dual- direction | 3 | 0.2413 | 0.4406 | 1.0053 | 0.2944 | 0.5219 | 1.1609 | 0.7373 | 0.1749 | -0.3083 |
| | 5 | 0.0738 | 0.5407 | 1.0043 | 0.0974 | 0.6101 | 1.2031 | 0.9715 | -0.1169 | -0.3344 |
| | 8 | 0.1031 | 0.3976 | 0.9307 | 0.1332 | 0.5097 | 1.0576 | 0.9169 | -0.2158 | -0.4234 |
| | 10 | 0.1404 | 0.5002 | 0.778 | 0.1691 | 0.5853 | 0.9284 | 0.8683 | -0.5786 | -0.2971 |

4. Conclusion

This paper proposed a novel method based on data fusion enhanced LSTM to estimate tool wear value under different cutting conditions. Firstly, the original vibration signals are decomposed and transformed to obtain high-dimensional feature series set through EMD, VMD and FSST, and then NCA is employed to select useful features and

reduce the feature dimension, in order to reduce operational burden and improve the accuracy of regression. Finally, these selected feature series are input into bidirectional LSTM network to estimate tool wear value. Hence, applications of the proposed method to milling TCM experiments demonstrate it outperforms significantly SVR- based and RNN- based methods under different cutting conditions.

References

- Cao XC, Chen BQ, Yao B, He WP. Combining translation-invariant wavelet frames and convolutional neural network for intelligent tool wear state identification. *Computers in Industry* 2019; 106: 71-84, <https://doi.org/10.1016/j.compind.2018.12.018>.
- Dong W, Tan X. Bayesian Neighborhood Component Analysis. *IEEE Transactions on Neural Networks & Learning Systems* 2018;29(7): 3140-3151, <https://doi.org/10.1109/TNNLS.2017.2712823>.
- Dragomiretskiy K, Zosso D. Variational mode decomposition. *IEEE Transactions on Signal Processing* 2014; 62(3): 531-544, <https://doi.org/10.1109/TSP.2013.2288675>.
- He K, Gao M, Zhao Z. Soft Computing Techniques for Surface Roughness Prediction in Hard Turning: A Literature Review. *IEEE Access* 2019; 7: 89556-89569, <https://doi.org/10.1109/ACCESS.2019.2926509>.
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation* 1997; 9(8): 1735-80, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hua YS, Mou LC, Zhu XX. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. *Isprs Journal of Photogrammetry And Remote Sensing* 2019; 149: 188-199, <https://doi.org/10.1016/j.isprsjprs.2019.01.015>.
- Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings Mathematical Physical & Engineering Sciences* 1998, 454(1971): 903-995, <https://doi.org/10.1098/rspa.1998.0193>.
- Huang Z, Zhu J, Lei J, Li X, Tian F. Tool wear predicting based on multi-domain feature fusion by deep convolutional neural network in milling operations. *Journal of Intelligent Manufacturing* 2020;31(4): 953-966, <https://doi.org/10.1007/s10845-019-01488-7>.
- Jasiulewicz-Kaczmarek M, Antosz K, Żywica P, Mazurkiewicz D, Sun B, Ren Y. Framework of machine criticality assessment with criteria interactions. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2021; 23(2): 207-220, <https://doi.org/10.17531/ein.2021.2.1>.
- Kozłowski E, Mazurkiewicz D, Zabinski T, Prucnal S, Sep J. Assessment model of cutting tool condition for real-time supervision system. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2019; 21(4): 679-685, <https://doi.org/10.17531/ein.2019.4.18>.
- Kozłowski E, Mazurkiewicz D, Żabiński T, Prucnal S, Sep J. Machining sensor data management for operation-level predictive model. *Expert Systems with Applications* 2020; 159: 1-22, <https://doi.org/10.1016/j.eswa.2020.113600>.
- Kumar A, Gandhi CP, Zhou YQ, Kumar R, Xiang JW. Improved CNN for the diagnosis of engine defects of 2-wheeler vehicle using wavelet synchro-squeezed transform (WSST). *Knowledge based System* 2020; 208, 106453, <https://doi.org/10.1016/j.knsys.2020.106453>.
- Kumar A, Gandhi CP, Zhou YQ, Kumar R, Xiang JW. Variational mode decomposition based symmetric single valued neutrosophic cross entropy measure for the identification of bearing defects in acentrifugal pump. *Applied Acoustics* 2020; 165, 107294, <https://doi.org/10.1016/j.apacoust.2020.107294>.
- Kumar A, Kumar R. Adaptive artificial intelligence for automatic identification of defect in the angular contact bearing. *Neural Computing & Applications* 2018; 29: 277-287, <https://doi.org/10.1007/s00521-017-3123-4>.
- Lei Z, Zhou YQ, Sun BT, Sun WF. An intrinsic time scale decomposition-based kernel extreme learning machine method to detect tool wear conditions in the milling process. *International Journal of Advanced Manufacturing Technology* 2020; 106(3-4): 1203-1212, <https://doi.org/10.1007/s00170-019-04689-9>.
- Lei Z, Zhu QS, Zhou YQ, Sun BT, Sun WF, Pan XM. A GAPSO-Enhanced Extreme Learning Machine Method for Tool Wear Estimation in Milling Processes Based on Vibration Signals. *International Journal of Precision Engineering and Manufacturing- Green Technology* 2021;

- 8: 745-759, <https://doi.org/10.1007/s40684-021-00353-4>.
17. Liu M, Wu W, Gu Z, Yu Z, Qi F, Li Y. Deep learning based on Batch Normalization for P300 signal detection. *Neurocomputing* 2018; 275: 288-297, <https://doi.org/10.1016/j.neucom.2017.08.039>.
18. Ma X, Tao Z, Wang Y, Yu H, Wang Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C- Emerging Technologies* 2015; 54: 187-197, <https://doi.org/10.1016/j.trc.2015.03.014>.
19. Musavi SH, Davoodi B, Eskandari B. Evaluation of surface roughness and optimization of cutting parameters in turning of AA2024 alloy under different cooling-lubrication conditions using RSM method. *Journal of Central South University* 2020, 27(6): 1714- 1728, <https://doi.org/10.1007/s11771-020-4402-2>.
20. Raghu S, Sriraam N. Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. *Expert Systems with Applications* 2018; 113: 18-32, <https://doi.org/10.1016/j.eswa.2018.06.031>.
21. Rosienkiewicz M. Artificial intelligence-based hybrid forecasting models for manufacturing systems. *Eksploracja i Niezawodność - Maintenance and Reliability* 2021; 23 (2): 263-277, <https://doi.org/10.17531/ein.2021.2.6>.
22. Sedlacek M, Podgornik B, Vizintin J. Influence of surface preparation on roughness parameters friction and wear. *Wear* 2009; 266(3-4): 482-487, <https://doi.org/10.1016/j.wear.2008.04.017>.
23. Siddhpura A, Paurobally R. A review of flankwear prediction methods for tool condition monitoring in a turning process. *International Journal of Advanced Manufacturing Technology* 2013; 65(1-4):371-393, <https://doi.org/10.1007/s00170-012-4177-1>.
24. Tao Z, An Q, Liu G, Chen M. A novel method for tool condition monitoring based on long short-term memory and hidden Markov model hybrid framework in high-speed milling Ti-6Al-4V. *International Journal of Advanced Manufacturing Technology* 2019; 105(7-8): 3165-3182, <https://doi.org/10.1007/s00170-019-04464-w>.
25. Tary JB, Herrera RH, Baan MV. Analysis of time-varying signals using continuous wavelet and synchrosqueezed transforms. *Philosophical Transactions of the Royal Society A* 2018; 376(2126), 20170254, <https://doi.org/10.1098/rsta.2017.0254>.
26. Tim VH, Chris KM. Self-supervised learning for tool wear monitoring with a disentangled-variational- autoencoder. *International Journal of Hydromechatronics* 2021; 4(1): 69-98, <https://doi.org/10.1504/IJHM.2021.114174>.
27. Wang B, Liu Z. Influences of tool structure tool material and tool wear on machined surface integrity during turning and milling of titanium and nickel alloys: a review. *International Journal of Advanced Manufacturing Technology* 2018; 98(5-8): 1925-1975, <https://doi.org/10.1007/s00170-018-2314-1>.
28. Zhang X, Zhao J. Compound fault detection in gearbox based on time synchronous resample and adaptive variational mode decomposition. *Eksploracja i Niezawodność - Maintenance and Reliability* 2020;22(1): 161-169, <http://dx.doi.org/10.17531/ein2020119>. <https://doi.org/10.17531/ein.2020.1.19>
29. Zhao R, Yan R, Wang J, Mao K. Learning to Monitor Machine Health with Convolutional Bi-Directional LSTM Networks. *Sensors* 2017; 17(2), 273, <https://doi.org/10.3390/s17020273>.
30. Zhi GF, He DD, Sun WF, Zhou YQ, Pan XM, Gao C. An edge-labeling graph neural network method for tool wear condition monitoring using wear image with small samples. *Measurement Science and Technology* 2021; 32(6), 064006, <https://doi.org/10.1088/1361-6501/abe0d9>.
31. Zhou JT, Zhao X, Gao J. Tool remaining useful life prediction method based on LSTM under variable working conditions. *International Journal of Advanced Manufacturing Technology* 2019; 104(9-12): 4715-4726, <https://doi.org/10.1007/s00170-019-04349-y>.
32. Zhou YQ, Sun BT, Sun WF. A tool condition monitoring method based on two-layer angle kernel extreme learning machine and binary differential evolution for milling. *Measurement* 2020; 166: 108186, <https://doi.org/10.1016/j.measurement.2020.108186>.
33. Zhou YQ, Sun BT, Sun WF, Lei Z. Tool wear condition monitoring based on a two-layer angle kernel extreme learning machine using sound sensor for milling process. *Journal of Intelligent Manufacturing* 2020, <https://doi.org/10.1007/s10845-020-01663-1>.
34. Zhou YQ, Xue W. Review of tool condition monitoring methods in milling processes. *International Journal of Advanced Manufacturing Technology* 2018; 96(5-8): 2509-2523, <https://doi.org/10.1007/s00170-018-1768-5>.
35. Zhu QS, Zhou YQ, Sun BT, He DD, Sun WF. A tool wear condition monitoring approach for end milling based on numerical simulation. *Eksploracja i Niezawodność - Maintenance and Reliability* 2021; 23(2):371-380, <https://doi.org/10.17531/ein.2021.2.17>.

A method for assessing of ship fuel system failures resulting from fuel changeover imposed by environmental requirements

Indexed by:



Przemysław Kowalak^{a,*}, Jarosław Myśków^a, Tomasz Tuński^a, Dariusz Bykowski^a, Tadeusz Borkowski^a

^aMaritime University of Szczecin, Mechanical Engineering Department, ul. Willowa 2, 71-650 Szczecin, Poland

Highlights


- Low sulfur fuels influence the failure rate of fuel injection system components.
- Failures were analyzed as related and not related to SECA fuel changeover.
- Quantitative and qualitative failure analysis was performed.
- Two markers were proposed for the quantitative assessment of failure frequency.
- Qualitative assessment was done by the adoption of three failure severity levels.

Abstract

Environmental regulations instigated the technological and procedural revolution in shipping. One of the challenges has been sulfur emission control areas (SECA) and requirement of fuel changeover. Initially, many reports anticipated that new grades of low sulfur fuels might increase various technical problems in ship operation. This research develops a simple and easy to use method of the failure severity and intensity assessment in relation to fuel changeover. The scale of failure rate in the ship's fuel system was evaluated qualitatively and quantitatively, using developed failure frequency indicator and the time between failure. Based on 77 records of fuel system failures collected on seven ships, it has been found that frequency of failures related to SECA fuel changeover is on average nearly three times higher compared to the rest of sailing time. Their severity did not significantly change, but the structure of failures changed considerably. The method and presented results may help in improvement of ship's systems design and on-board operational procedures.

Keywords

failure frequency, fuel changeover, fuel system failure, emission control area.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

1. Introduction

Established by Annex VI to MARPOL 73/78 Convention [18], sulfur emission control areas (SECA) followed by the global stepwise limitation of sulfur content in marine fuels, resulted in several changes in ship construction, performance and operation. Fuel oil bunkering and storage systems had to be redesigned and some fuel oil tanks had to be designated for low sulfur fuel oils (LSFO) storage [6, 12, 23, 34]. In many cases, additional cylinder oil storage and supply systems had to be provided to allow smooth and safe fuel changeover [24–26]. Additionally, the engine cylinder components, like pistons and piston rings, had to be modified [27] to improve the engine reliability and sustainability when operating on fuel grades different from the design. An example of such modifications is the high temperature cylinder cooling system, which was introduced on very long stroke engines around the year 2014 to counteract a low temperature corrosion, and was later recommended to be deactivated for engines enduringly consuming fuel oil with 0.5% of sulfur or less [28]. All those examples show the difficulties and complexity of problems related to low sulfur fuels faced by equipment makers, shipowners and, finally, the crew. Crews in particular are burdened with additional maintenance and adjusting work, and in case of machinery failure, with extra service work [13].

After the first SECA, covering initially the Baltic Sea and the North Sea, was established, soon other were implemented in various regions of the world, starting from the North America and some regions of the Caribbean Sea, to a number of Chinese ports. That resulted in frequent fuel changeovers from high sulfur fuels used in the open ocean passage to low sulfur residual fuels or even low sulfur distillate fuels [24]. The past few years have witnessed the introduction of another fuel grade called ultra-low sulfur fuel (ULSFO), or hybrid fuel, being a stabilized blend of very low sulfur distillate fuel with residual fuel.

Low sulfur fuel oils have different properties, especially viscosity, stability and lubricity, compared to typical high sulfur fuel oils [1, 12, 34]. Viscosity, which is directly dependent on fuel temperature, may play a key role in failures of fuel injection system components [5, 6, 26]. Most of the engine arrangements could not be quickly adapted to the use of low sulfur fuels [2, 13, 29]. Consequently, ship operators faced significant problems with machinery operation and the number of reported incidents related to fuel changeover raised significantly [3, 13, 17].

To safely perform a fuel changeover, shipping companies and ship's crew developed and implemented new procedures [19, 24]. The time required for the proper and safe procedure depends mainly on the sulfur content in the high and low sulfur fuels to be altered, the en-

(*) Corresponding author.

E-mail addresses: P. Kowalak - p.kowalak@am.szczecin.pl, J. Myśków - j.myskow@am.szczecin.pl, T. Tuński - t.tunski@am.szczecin.pl, D. Bykowski - d.bykowski@am.szczecin.pl, T. Borkowski - t.borkowski@am.szczecin.pl

gine load, and finally the fuel system volume to be flushed [6, 13, 23, 24, 26]. At least one of those parameters, namely, the engine load, is variable and depends greatly on the weather and nautical conditions, consequently, the entire procedure may take from a few hours up to two or three days even. The initial phase of changeover is crucial for the machinery and consequently ship safety. It has to be carried out slowly and with utmost care to avoid rapid changes in fuel temperature and viscosity [13, 23, 26, 36]. During this phase, the fuel pipes trace heating has to be stopped, the fuel viscosity controllers usually have to be set to manual mode and the fuel temperature gradually reduced to maintain a safe fuel lubricity level. One of the frequent problems is the deposit formulation during mixing of different fuel grades boosted by altering temperature, which leads to clogging of filters and disturbances in viscosimeter readings [21].

Even so, it is observed that despite the utmost care during fuel changeover, ship fuel systems suffer an increased number of incidents related to the malfunction of equipment, chiefly filters, centrifuges, heaters, and engine fuel injection components.

Legislative bodies, such as the International Maritime Organization (IMO), the European Union (EU), or port state authorities, impose increasingly stricter environmental requirements on sea-going ships. The necessity to reduce environmental pollution is beyond dispute. However, no means exist to verify the impact of the applied legal requirements on the technical condition of ships, their safety and reliability.

Increasingly complex and demanding devices to reduce the emission of harmful substances into the environment are being installed on ships. Ship crews are burdened with additional duties related to their operations and new environmental procedures. Shipowners do not have any incentives to increase the number of crews beyond the safety regulations and economic demand. Consequently, the risk of machinery failure or an accident may rise [6].

Although legislators make efforts to monitor and control the process of adopting new regulations, the main focus is on compliance verification and the influence on economy. For example, most classification societies issued dedicated fuel changeover guidelines for shipping companies and ships' staff [2, 12]. The European Maritime Safety Agency (EMSA) regularly issues updates to sulfur inspection guidance [14]. Problems widely analyzed by the states are the economic impact and low sulfur fuels availability. A number of related publications and reports were issued over the last decade [4, 8, 10, 15, 16, 31]. However, there are few reports or research publications analyzing the problem of machinery reliability and failure intensity related to the fuel changeover procedures. Statistics published by the French Ministry of Environment revealed that in 2015 the number of reported loss of power incidents in the English Channel doubled compared to the previous year [17]. The positions of ships report-

ing incidents suggest that they may be related to fuel changeover on entering or leaving SECA. Very similar increase in the loss of power was observed in 2019 in California after the California Air Resources Board regulation entered into force [17]. Some accidents, especially those leading to injuries or severe loss in property or environment, are reported to authorities and after investigation reports are made public [37]. However, information is scarce about the number of failures which were not officially reported. Is there a similar rate of failures compared to the pre-SECA conditions? Regulators, interested in meeting the requirements by ships, should also have ship safety and reliability in focus. Proper feedback may and should be taken into account when a regulation is revised or updated, and/or guidance for procedures is being prepared. However, the record of officially reported accidents may be insufficient. There are multiple cases of different malfunctions and incidents that have never been reported to organizations other than the shipowner's company, while each such case may trigger a chain of events leading to disaster. The newly introduced regulations have their consequences: those expected, but also unexpected side effects. Assessing possible negative consequences may play a key role in improving ship safety.

In this research, the frequency of failures and malfunctions in the ship fuel systems related to fuel changeover, including engines, supply, and injection system, was analyzed and compared to the frequency of similar incidents occurred during engines operation on one grade of fuel only.

2. Analysis object and method

Statistical data were collected on seven merchant ships of various types and capacity: four container carriers and three multipurpose general cargo vessels (Table 1). During the period of observation, four ships were not older than three years, while the remaining three ships were 8 to 10 years old. The selected ships entered a SECA at least once during the observation period. Because the trading areas cover almost all the oceans and to simplify the nomenclature, the SECA in this research means all areas where the limits of sulfur content in marine fuels were imposed, especially: Northern Europe, North American coast, Caribbean Sea region and Chinese Pearl River Delta, Yangtze River Delta and Bohai Bay. The deck and the engine logbooks of each ship were analyzed to determine the exact time of fuel changeover commencement and completion when entering and leaving SECA.

Due to the relatively long observation period, starting in 2010 for ship A and ending in 2020 for ship G, the requirements for SECA differ depending on the actual date and port of call. Consequently, the fuel grades used on board the selected ships also differed according to the evolution of sulfur limits inside and outside SECAs (Table 1).

Table 1. Basics of the analyzed ships and their voyages

| Ship | Year and place of build | DWT, tons | Propulsion type | Period of observation | Fuel grades used | Number of SECA calls |
|------|-------------------------|-----------|-----------------|-----------------------|-------------------|----------------------|
| A | 2010, China | 50300 | Direct, FPP | Jan 2010 – Sep 2012 | HSHFO/LSHFO/LSMGO | 7 |
| B | 2014, China | 60550 | Direct, FPP | Jun 2014 – Oct 2014 | HSHFO/LSHFO/LSMGO | 4 |
| C | 2012, South Korea | 145451 | Direct, FPP | Mar 2015 – Jul 2015 | HSHFO/LSMGO | 2 |
| D | 2014, South Korea | 149360 | Direct, FPP | May 2015 – Jul 017 | HSHFO/LSMGO | 38 |
| E | 2009, China | 7811 | Indirect, CPP | Apr 2017 – Jul 2017 | HSHFO/LSMGO | 5 |
| F | 2011, China | 5646 | Indirect, CPP | May 2018 – Apr 2019 | HSHFO/LSMGO | 6 |
| G | 2010, China | 12940 | Indirect, CPP | Mar 2020 – Jul 2020 | LSMGO/ULSHFO | 4 |

DWT – deadweight tonnage
HSHFO – high sulfur residual fuel (as defined in the regulations currently in force),
LSHFO – low sulfur residual fuel,
LSMGO – low sulfur marine gasoil
ULSHFO – ultra low sulfur residual fuel (hybrid fuel)
CPP – controllable pitch propeller
FPP – fixed pitch propeller

All selected ships, except ship A, were calling SECA regularly. The trading area of ship A was outside the SECA for the first two years of the analyzed period, followed by a series of voyages between Central America and the European SECA in 2012, therefore most of identified on this ship failures is not related to SECA fuel changeover.

Because the fuel system malfunction may occur with some delay after the fuel changeover procedure is accomplished, it was arbitrarily decided that the incident is related to fuel changeover if it occurs after the commencement of the procedure, not later than three days after its completion. All other incidents are assumed as not directly related to the fuel changeover procedure. With that assumption, callings at SECA lasting over six days were assigned six to seven days of observation per each calling, depending on the time required for completion of the changeover procedure. That was frequent case for ships calling at ports situated in the North Sea and Baltic Sea SECA region where typically more than one port were visited and the entire sea passage between them is within a single SECA. On the other hand, in case of short calls, less than three days in SECA, the time of observation was three to six days depending on the length of berthing time. This applied particularly to calls at a single port in North America, or, since January 2016, at Chinese ports in Pearl River Delta, Yangtze River Delta or Bohai Bay.

Based on the deck and the engine logbooks entries we determined the time of the fuel changeover observation T_{CO} and calculated the ratio R_o of observation time between the T_{CO} and the total observation time T_{tot} :

$$R_o = \frac{T_{CO}}{T_{tot}} \cdot 100\% \quad (1)$$

where: T_{CO} – time of the fuel changeover observation; T_{tot} – total observation time; R_o – ratio of observation time.

Similarly, the engine logbooks and other official reporting documents, like near miss reports, malfunction reports, damage reports and repair reports were analyzed for evidence of incidents related to ship fuel system failures. Identification of historical failures was frequently facilitated by ship's photo documentation, where an actual date of the failure was usually recorded. We also used monthly work reports – internal reports of the shipping companies. All identified failures were assigned the date and if possible, the time of occurrence. The study covered the entire fuel system: storage, transfer, purification, supply to the main engine, auxiliary engines and fired boiler, and finally the engine injection system. All routine service and maintenance work, such as time-based fuel injection pumps or fuel injection valves maintenance, was excluded from the analysis.

The proposed analysis makes use of some elements and techniques adopted from the reliability engineering [22, 32, 33], mainly Failure Mode and Effect Analysis (FMEA). Because of the varying nature and location of failures, it is practicable to group them with respect to the most relevant parameter [33, 38]. A similar method was applied in this study and the identified failures were classified into three classes of location:

1. Class A. Failure of the engine fuel injection system. This group includes malfunction of fuel injection valves (FIV), fuel injection pumps (FIP), high pressure injection pipes, and their safety system – leakage detection system for both main engine and auxiliary engines.
2. Class B. Failure of the fuel supply system. This group includes fuel supply and circulation pumps, fuel safety filters, fuel automatic filters, fuel preheaters and coolers, viscosity sensors, fuel supply pipes, and their tracing heating.
3. Class C. Failure of the fuel storage, transfer, and preparation system, including the purification system. This group includes mainly problems in storage, settling, or service tanks (sediments, contamination, foaming), difficulties with transporta-

tion related to fuel properties, contaminated filters, strainers, or purifiers and their preheaters.

The definition of failure is always problematic and a variety of approaches are proposed by different researchers [9, 20, 32, 35]. In essence, based on the ISO 8402 the definition of reliability [32], failure may be defined as the inability to perform a required function under given environmental and operational conditions and for a stated period of time. However, in ship service, situations occur where a component or subsystem is functioning, but the risk of accident or loss of property is very high. Such a situation is called a near miss incident. Therefore, for this research, a total inability to perform a function, as well as a near miss condition and malfunctions likely leading to a near miss are recognized as failures. Similar approach is described in the literature [7, 11].

For every recorded failure, the severity of its actual or possible consequences was evaluated too. Again, similarly to the definition of failure, there is no single universal definition of severity levels. For example, Morais [30] proposes a very simple classification into three levels of severity: no problem, moderate problems, and extreme problems, which seems to be very universal and applicable in various disciplines. However, in case of failure consequences analysis, the lowest of proposed levels may be inadequate. A more suitable definition was proposed by Kaidis [20], who related the severity levels to the required service time. Sasmito and Untung proposed a criticality of failure matrix with four categories of failure severity for the analyzed ship's fuel system [33]. In fact, severity should be individually defined to the needs of the specific problem. Therefore, in this work three levels of severity were defined:

1. High risk failure – when the vessel had to be stopped, departure was delayed or an auxiliary engine or fired boiler could not be started for at least one hour.
2. Medium risk failure – when the ship operation was not disturbed, but there was a direct and significant risk of disturbance leading to a high-risk incident, similar to a near miss condition.
3. Low risk failure – when the ship operation was not disturbed and there was no direct and significant risk of disturbance leading to a high-risk incident.

Of all identified failures, those related to the fuel changeover procedure were selected based on the date and time of occurrence. Additionally, they were evaluated by an experienced engineer on board the ship for possible relation to fuel changeover procedure. Even if it is unavoidable to have such evaluation biased by an individual and subjective judgment, the authors chose to do so as the risk of erratic qualification was thought to be lower when engineer's evaluation is done than when it is not. Finally, the number of failures related to fuel changeover n_{CO} and the total number of failures n_{tot} were used to calculate the ratio of failure occurrence R_{foc} for every individual ship and for the whole analyzed population:

$$R_{foc} = \frac{n_{CO}}{n_{tot}} \cdot 100\% \quad (2)$$

where: n_{CO} – number of failures related to fuel changeover observed during the time T_{CO} ; n_{tot} – total number of failures observed during the time T_{tot} .

Dividing the ratio of failure occurrence R_{foc} by the ratio of observation time R_o , we can determine the failure frequency indicator F_i :

$$F_i = \frac{R_{foc}}{R_o} \quad (3)$$

The failure frequency indicator F_i should be close to unity if the frequency of failures related to fuel changeover in SECA and the

overall failures frequency are similar. In case failures related to fuel changeover in SECA are more frequent, the value of F_i rises above unity. That makes the F_i very easy to interpret.

Additionally, the time between failures (TBF) was calculated for each class of failure and each ship using the formula:

$$TBF_{class, condition} = \frac{T_{condition}}{n_{class, condition}} \quad (4)$$

where: *class* – is the location of failure according to the presented classification A, B, C; *condition* – is the condition of observation: related to SECA fuel changeover or not related to fuel changeover; $TBF_{class, condition}$ – time between failure of a specific class in a specific condition [days]; $T_{condition}$ – time of observation [days]; $n_{class, condition}$ – number of incidents of a specific class and in specific conditions.

For calculation of TBF related to SECA fuel changeover, T_{CO} was used in the formula (4) numerator, while to calculate TBF not related to SECA fuel changeover, the difference $T_{tot} - T_{CO}$ was applied. This approach is different from the way the failure frequency indicator F_i is calculated, for which the time of the fuel changeover observation T_{CO} is divided by the total observation time T_{tot} instead of the difference $T_{tot} - T_{CO}$. That is mainly to bring the formula (4) as close as possible to the way the MTBF (mean time between failures) is calculated in the theory of reliability. However, the above defined TBF should not be understood as a typical MTBF. It is rather a quantitative estimation of the likelihood of a specific malfunction in specific conditions. By definition, the MTBF is calculated from the working time of the component, while in this study the TBF was evaluated from the failure-to-failure time span regardless of whether the component was running or stopped during that time. Moreover, the limited statistical sample makes the generalized result very uncertain to use the term MTBF.

3. Analysis of failure structure

77 failures were identified on all seven ships during the total observation time. Only one of them was officially reported to a Vessel Traffic Service (VTS) on the French coast, while the remaining 76 failures were just recorded in the ship's documentation; only 40 of them were also reported to the owner's office. The remaining 36 failures were only noted in the ship's documentation without any official reporting. The number of minor failures without sufficient documentation is not known, although evidence was found, like improperly described photos, that such failures also had occurred.

The structure of failures with respect to the affected component is presented in Table 2. The component with the highest number of recorded failures in class a is the fuel injection valve (FIV) with the

total 16 cases. The fuel injection pump (FIP) ranks second with 12 cases of failure recorded.

For the analyzed population of ships, there is no difference observed in the severity of FIV failure between related and not related to SECA fuel changeover (Fig. 1). However, it should be noticed, that the number of analyzed failures is only 16. It is very likely, that longer observation time or larger population of ships could reveal some differences.

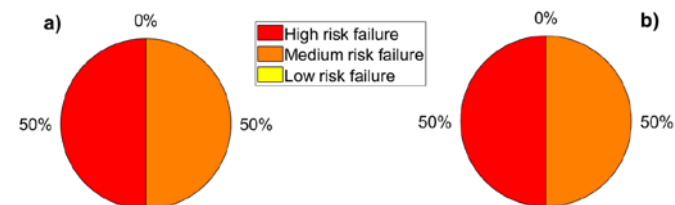


Fig. 1. Comparison of the FIV failures structure with respect to the failure severity: a) failures not related to SECA fuel changeover; b) failures related to SECA fuel changeover

It is symptomatic that due to the function of FIV, there are no low-risk failures observed at all. Once the FIV performance is deteriorated, it usually requires urgent or even immediate action. In most cases of high-risk failures, severe mechanical destruction of the FIV is observed, frequently accompanied by fuel leakage into the engine combustion chamber. Figure 2 depicts two different cases of two-stroke engine FIV with broken nozzle tips. The left-hand photo presents damage not related to SECA fuel changeover, while the damage presented in the right-hand photo was observed 20 hours after the fuel changeover procedure commencement. In both cases, the engine had to be stopped for FIV replacement.

An additional example is shown in Fig. 3, where the damaged FIV suffered a strong impact of exhaust gas blow-by through the seating. The failure occurred six hours after changeover from residual to low sulfur distillate fuel commencement while entering the European SECA. This specific incident resulted in damaged engine cylinder cover, temporary cut-out of the failed engine cylinder, and emergency steaming to the port of destination. That was the only officially reported incident in the entire analyzed population.

Generally, the most severe failure of FIP is the seizing of the plunger and barrel. It is nearly always qualified as a high-risk failure as it usually requires engine shutdown. It may be caused by inadequate fuel purifying or filtering. It also frequently happens as a result of a low viscosity and lubricity of the fuel, especially when the introduced distilled fuel has a low sulfur content or experiences a drastic decrease

Table 2. Comparison of the number of failures for classes of location A, B, C with respect to the affected components

| Affected component | Number of failures not related to SECA fuel changeover | | | Number of failures related to SECA fuel changeover | | |
|---------------------|--|---------|---------|--|---------|---------|
| | Class A | Class B | Class C | Class A | Class B | Class C |
| FIP | 7 | - | - | 5 | - | - |
| FIV | 10 | - | - | 6 | - | - |
| HP pipes | 3 | - | - | - | - | - |
| Return/supply pipes | 5 | 3 | - | - | 3 | - |
| Pumps | - | 5 | - | - | 1 | - |
| Filters | - | 8 | 1 | - | 7 | - |
| Tank contamination | - | - | 2 | - | - | 1 |
| Purifiers | - | - | 1 | - | - | 4 |
| Tank structure | - | - | 1 | - | - | - |
| Heating and tracing | - | - | 2 | - | - | - |
| Tank level sensor | - | - | 2 | - | - | - |



Fig. 2. FIV nozzle tip damages assigned to the high risk failure group: left-hand photo – failure not related to SECA fuel changeover; right-hand photo – failure related to SECA fuel changeover



Fig. 3. High risk damage of FIV related to SECA fuel changeover, i.e. seating burnt out by combustion gas blow-by

in viscosity due to excessively high temperature. This effect can be significantly accelerated by a large amount of heat accumulated in the elements of injection pumps during the changeover from residual to distillate fuels. Most of the engines accept the distillate fuel kinematic viscosity not lower than 2-3 mm²/s, which means that the temperature of the distillate fuel supplied to the engine should be maintained below 50°C. But the temperature of the residual fuel frequently exceeds 140°C. Consequently, during changeover the viscosity of the distillate fuel may drop below that recommended by the engine maker. Even more problematic is the changeover from distillate to residual fuels. If the warm-up process is too fast, the plunger expands faster than the barrel, causing a dangerous decrease of a very fine clearance required for movability of the elements, frequently resulting in seizures [26].

For the analyzed population of ships, most FIP failures were qualified as high or medium risk (Fig. 4), but the share of high-risk failures requiring immediate engine shutdown raised from 29% to 50% in relation to SECA fuel changeover. An example of a FIP plunger damage occurred during rapid fuel changeover is shown in Fig. 5.

Observed medium risk failures were usually FIP non-return valve malfunctions or moderate fuel leaks. In one case it was short stuck of the plunger and barrel which became movable after a few seconds. The FIP was replaced in the next port, a few days after the incident.

The only case of low-risk failure observed in a group of failures not related to SECA fuel changeover (Fig. 4a) was a fuel leakage through an internal seal resulting in minor lubricating oil contamination.

Other components of the fuel system with a sufficient number of recorded failures are the filters in the supply system of failure class b. Surprisingly, in the analyzed population of ships, the severity structure of filter failures due to SECA fuel changeover or other causes is much different than expected. Seafarers, when interviewed, tend to complain about the incompatibility of different fuels grades and frequent problems with filter clogging, formation of sediments, and extreme gasification. The graphs presented in Fig. 6 do not confirm that the severity of those problems is greater when fuel is changed over in SECA compared to the severity of similar incidents during changeover of fuel not related to SECA. However, the frequency of problems with proper filtration is still higher

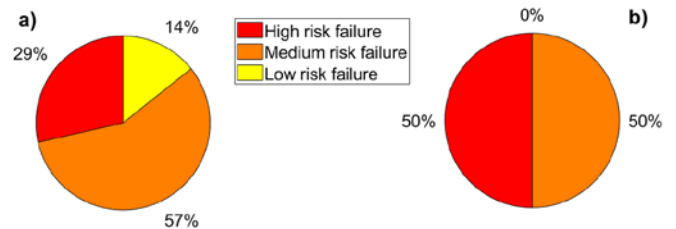


Fig. 4. Comparison of the FIP failures structure with respect to the failure severity: a) failures not related to SECA fuel changeover; b) failures related to SECA fuel changeover



Fig. 5. Fuel injection pump plunger seizure occurred during rapid fuel changeover from residual to low sulfur distillate fuel

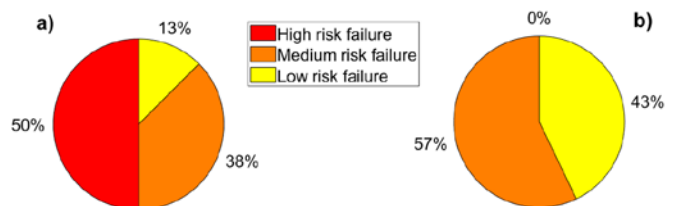


Fig. 6. Comparison of the fuel filter failures structure with respect to the failure severity: a) failures not related to SECA fuel changeover; b) failures related to SECA fuel changeover

in SECA related group. It is possible that the ship crew is much more careful and prepared for possible problems when fuel changeover is carried out in SECA, which results in the actual elimination of high-risk failures. Nevertheless, in the proposed analysis this hypothesis has not been verified.

4. Results and discussion

The total observation time T_{tot} of the selected population of ships was 2652 days. During this time the analyzed ships entered SECA with various frequencies, and the time spent in SECA differed from single days to a week or more. Consequently, the individual ship observation time ratio R_o calculated by formula (1) varied from 4.4% to 19.8%. The observation time ratio was also calculated for the entire population of the analyzed ships:

$$R_{ofleet} = \frac{\sum_i T_{COi}}{\sum_i T_{toti}} \cdot 100\% \quad (5)$$

where: T_{COi} – time of the ship i fuel changeover observation; T_{toti} – total observation time of the ship i .

The overall span of R_o is only 15.4 %, with the fleet observation time ratio $R_{ofleet}=11.9\%$ (Table 3), which indicates that no extreme differences existed between ships in the intensity of callings at SECA.

After thorough verification of available documentation, 27 failures were qualified as failures related to fuel changeover during entering or leaving SECA. Failures occurred during routine changeover of the

Table 3. Comparison of the total observation time T_{tot} and time of changeover observation T_{CO} for the analyzed ships

| Ship | T_{tot} , day | T_{CO} , day | R_o , % |
|--------------|-----------------|----------------|-----------|
| A | 958 | 42 | 4.4 |
| B | 127 | 22 | 17.3 |
| C | 141 | 12 | 8.5 |
| D | 801 | 158 | 19.7 |
| E | 111 | 22 | 19.8 |
| F | 357 | 36 | 10.1 |
| G | 157 | 24 | 15.3 |
| Total | 2652 | 316 | - |
| R_{ofleet} | | | 11.9 |

same grades of fuels from different bunker suppliers, but those not related to entry or leaving from SECA were not assigned to this group. The number of failures, divided into three classes: a, b or c, and into groups of related and not related to SECA fuel changeover, are presented in Table 4.

Based on the number of failures identified for each ship (Table 4), the ratio of failure occurrence R_{foc} was calculated with formula (2). Similarly to the observation time ratio, the results varied, but the span was much wider: from 12% to 83.3% (Table 4). For each ship except ship E, the values of R_{foc} are significantly higher than R_o . The average R_{foc} for all ships (35.1%) is nearly three times higher than the overall average of R_{ofleet} (11.9%). This indicates that for the analyzed population of ships, failures in the fuel system were observed on average three times more frequently during fuel changeover in SECA compared to the total average frequency.

Table 5. Comparison of average TBF for classes of location A, B, C of failure related and not related to SECA fuel changeover

| Ship | TBF not related to SECA fuel changeover, day | | | TBF related to SECA fuel changeover, day | | |
|---------|--|---------|---------|--|---------|---------|
| | Class A | Class B | Class C | Class A | Class B | Class C |
| A | 65 | 305 | 183 | 21 | 42 | - |
| B | 105 | 105 | - | 11 | 22 | - |
| C | 43 | 129 | - | - | 4 | - |
| D | 92 | 129 | 322 | 40 | 53 | 40 |
| E | - | 30 | 89 | - | 22 | - |
| F | - | 161 | 321 | 36 | - | - |
| G | - | 133 | - | 12 | 12 | 24 |
| average | 76 | 142 | 229 | 24 | 26 | 32 |

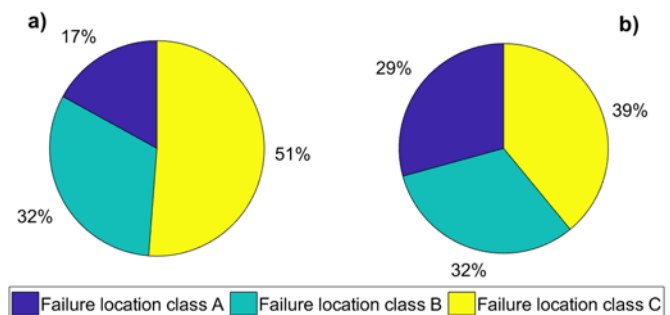


Fig. 7. Comparison of the failures structure with respect to the failure location class: a) failures not related to SECA fuel changeover; b) failures related to SECA fuel changeover.

Table 4. Number of failures related and not related to fuel changeover in SECA

| Ship | Number of failures not related to SECA fuel changeover ($n_{tot} - n_{CO}$) | | | Number of failures related to SECA fuel changeover n_{CO} | | | Ratio of failure occurrence R_{foc} , % | Failure frequency indicator F_i |
|--|---|---------|---------|---|---------|---------|---|-----------------------------------|
| | Class A | Class B | Class C | Class A | Class B | Class C | | |
| A | 14 | 3 | 5 | 2 | 1 | - | 12.0 | 2.7 |
| B | 1 | 1 | - | 2 | 1 | - | 60.0 | 3.5 |
| C | 3 | 1 | - | - | 3 | - | 42.9 | 5.0 |
| D | 7 | 5 | 2 | 4 | 3 | 4 | 44.0 | 2.2 |
| E | - | 3 | 1 | - | 1 | - | 20.0 | 1.0 |
| F | - | 2 | 1 | 1 | - | - | 25.0 | 2.5 |
| G | - | 1 | - | 2 | 2 | 1 | 83.3 | 5.4 |
| total | 25 | 16 | 9 | 11 | 11 | 5 | - | - |
| Average value for entire population of ships | | | | | | | 35.1 | 2.9 |

The TBF calculated with formula (4) and presented in Table 5 is even better indicator of SECA fuel changeover influence on the machinery reliability. The average TBF related to SECA fuel changeover is three to seven times shorter for each class of location: A, B, and C, compared to the TBF not related to SECA fuel changeover. Moreover, for every individual ship and class of location, TBF related to SECA fuel changeover is shorter. The structure of the failures is different, too (Fig. 7). The share of the fuel injection system failures (failure class a) increased from 17% of the total not related to SECA fuel changeover cases to 29% of related to SECA fuel changeover cases. While the share of failure class b of the fuel supply system remains unchanged (32%), the share of failures in the fuel storage and preparation system dropped when fuel is changed over in SECA from the initial 51% to 39%. The presented results suggest that the fuel changeover in SECA affects the injection system rather than the fuel storage and preparation system. However, the problems in the latter system are likely to occur prior to the actual commencement of the fuel changeover procedure, mostly due to the necessity to commence new fuel preparation well in advance: preheating, purifying, and transfer. The method used in this research does not allow confirming this hypothesis and should be verified in a separate research.

6. Conclusion

The presented analysis is aimed at emphasizing the problem of technical consequences related to the changeover to low sulfur fuel while entering or leaving SECA. The population of analyzed ships is not numerous to draw a generalized conclusion for the larger fleet. However, even for a small sample, differences are observed between the failure frequencies and time between failures of specific components. The presented analysis results and the method of data processing is a proposal highlighting the fuel oil changeover problem rather than a general recommendation.

The proposed method of analysis allows for both quantitative and qualitative assessment. There are two indicators proposed for the quantitative assessment of failure frequency. The failure frequency indicator F_i allows us to assess promptly and easily whether the failures occur more or less frequently in relation to SECA fuel changeover. For all the examined ships, the individual F_i is greater than 1. The average for the entire population is $F_i=2.9$ (Table 4), which suggests that the likelihood of failure in the fuel oil system is on average nearly three times higher while entering or leaving SECA compared to the entire operation time of all analyzed ships. Presented in Table 5, the values of time between failure TBF correspond with F_i . In the group of failures related to SECA fuel changeover, the average TBF is 24, 26, and 32 days for the respective failure location class A, B and C, compared to TBF not related to SECA fuel changeover, 76, 142 and 229 days, respectively. It means that in the analyzed population of ships, the TBF related to SECA fuel changeover is threefold shorter in the failure location class A, over fivefold less in the failure location class B, and seven times shorter in the failure location class C.

The qualitative assessment was achieved by the adoption of the failure severity metrics, where three levels of severity were defined: low, medium, and high. While failures of nearly all analyzed components in all classes are observed much more frequently when the

ship enters or leaves SECA compared to the frequency of failures not related to SECA fuel changeover, the observed severity of failures is not necessarily increased in relation to SECA fuel changeover. Due to the limited amount of data, only failures of three components: FIP and FIV of failure location class A, and fuel filters of failure location class B were analyzed qualitatively. Only in case of FIP, the share of high-risk failures grew from 29% to 50% with a simultaneous decrease of low-risk failures from 14% to 0%. For the remaining two components, namely FIV and filters, no increase in failure severity was observed. The presented qualitative results, due to the relatively small samples of the input data, show only the feasibility of the analysis rather than the overall conclusion for the larger fleet.

In the proposed method, most data were derived from the ship's internal records. Only one out of 77 failures qualified in the research were officially reported to the authorities, which shows the scale of unknown technical problems faced by the ships and their crews. It also proves that there is a space for improvement in terms of technical monitoring procedures.

In this research, the fuel system was chosen as an example. However, there are also other systems and machinery on board the ship which may be affected by the fuel changeover, like exhaust gas system, heating system, boilers, main and auxiliary engines. It might be especially important to establish how the specific low sulfur fuel grades influence the machinery reliability during changeover. Unfortunately, the insufficient population of seven ships prevents effective analysis. The proposed method is very flexible and may be easily adapted to the specific needs of any ship system or machinery and to any existing or future regulatory requirements.

Even if the applied methods are very simple, they proved to be effective: similar methods are used in industrial reliability analyses. The simplicity is a great advantage in this case. Availability of source data should not pose any difficulty, the utilized data are relatively easy to access on every sea-going ship, so what remains is standardized processing. Moreover, the crew engaged in data collection should not be burdened with additional work, provided a standardized and anonymized system of reporting failures, damage and incidents is introduced. Such a uniform system would probably significantly facilitate the process for crews by elimination the need to learn new procedures of reporting when changing the shipowner.

In the Authors opinion, a similar approach might be a good tool for a large-scale analysis. Information derived may be useful for fleet operators, the authorities and legislators, and especially for ships and machinery designers. Proper cooperation of ship operators, designers, shipbuilders, policymakers, authorities and ship personnel is crucial for effective and safe introduction of new environmental policies.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References:

1. Alfa Laval. Alfa Laval - Marine fuels in the low-sulphur era. Lund 2018. [<https://www.alfalaval.com/industries/marine-transportation/marine/oil-treatment/fuel-line/marine-fuels-in-the-low-sulphur-era/>].
2. American Bureau of Shipping. Fuel Switching Advisory. Houston 2015. [https://ww2.eagle.org/content/dam/eagle/advisories-and-ebriefs/ABS_Fuel_Switching_Advisory_15076.pdf].
3. Anh Tran T. Some Methods to Prevent the Wear of Piston-Cylinder When Using Low Sulphur Fuel Oil (LSFO) for All Ships Sailing on Emission Control Areas (ECAs). Diesel and Gasoline Engines, 2020. <https://doi.org/10.5772/intechopen.89400>.
4. Antturi J, Hänninen O, Jalkanen J P et al. Costs and benefits of low-sulphur fuel standard for Baltic Sea shipping. Journal of Environmental Management 2016; 184: 431–440, <https://doi.org/10.1016/j.jenvman.2016.09.064>.
5. Bejger A, Drzewieniecki J. Analysis of tribological processes occurring in precision pairs based on example of fuel injection pumps of marine

- diesel engines. *Scientific Journals of the Maritime University of Szczecin* 2015; nr 41(113): 9–16.
6. Borkowski T, John A. State of Play and Future needs for Clean Shipping. Szczecin and Rostock, 2021. [<https://cshipp.eu/wp-content/uploads/2021/03/State-of-Play-and-Future-Needs-for-Clean-Shipping-report.pdf>].
 7. Chen H, Moan T. Collision Risk Analysis of FPSO-Tanker Offloading Operation. *Proceedings of the 21st International Conference on Offshore Mechanics and Arctic Engineering* 2002; 2: 101–112, <https://doi.org/10.1115/OMAE2002-28103>.
 8. Chu Van T, Ramirez J, Rainey T et al. Global impacts of recent IMO regulations on marine fuel oil refining processes and ship emissions. *Transportation Research Part D: Transport and Environment* 2019; 70: 123–134, <https://doi.org/10.1016/j.trd.2019.04.001>.
 9. Chybowski L, Gawdzińska K, Laskowski R. Assessing the unreliability of systems during the early operation period of a ship-A case study. *Journal of Marine Science and Engineering* 2019; <https://doi.org/10.3390/jmse7070213>.
 10. Czermański E, Drożdziejcki S, Matczak M et al. Sulphur Regulation-Technology Solutions and Economic Consequences for the Baltic Sea Region Shipping Market. Institute of Maritime Transport and Seaborne Trade University of Gdańsk: 2014 .
 11. Department of Natural Resources Mines and Energy. Guidelines for Failure Impact Assessment of Water Dams. 2018. [https://www.dews.qld.gov.au/_data/assets/pdf_file/0005/78836/guidelines-failure-impact-assessment.pdf].
 12. DNV-GL. Technical Update - Preparing For Low Sulphur Operations. Hamburg 2014. [<https://margetis.com/wp-content/uploads/2019/01/DNV-GL-Technical-Update-preparing-for-low-sulphur.pdf>].
 13. ECSA. Overview of ‘fuel changeover’ issues and challenges as they affect ECA SOx compliance. 2014. [<https://www.ecsa.eu/sites/default/files/publications/C-8690%20Annex%201%20-%202014-11%20fuel%20changeover%20%20ics%20ecsa.pdf>].
 14. European Maritime Safety Agency (EMSA). Sulphur Inspection Guidance Directive (EU) 2016/802. 2019. [<http://www.emsa.europa.eu/publications/reports/item/2407-sulphur-inspection-guidance.html>].
 15. European Maritime Safety Agency (EMSA). The 0.1% sulphur in fuel requirement as from 1 January 2015 in SECAs. 2010. [https://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKewi-7evTwtjxAhXmmIsKHRE8AbUQFjACegQICBAD&url=https%3A%2F%2Fwww.nepia.com%2Fmedia%2F221111%2FReport_Sulphur_Requirementpdf_c_.pdf&usq=AOvVaw3qyPZWCYcQcuxvMhonsu7i].
 16. Fan L, Gu B. Impacts of the increasingly strict sulfur limit on compliance option choices: The case study of Chinese SECA. *Sustainability (Switzerland)* 2020; <https://doi.org/10.3390/SU12010165>.
 17. Ian Crutchley. Adjusting to change. *Bunkerspot* 2016; 68–70. [<https://innospec.com/wp-content/uploads/2020/10/Adjusting-to-change-Bunkerspot-August-2016.pdf>].
 18. IMO. International Convention for the Prevention of Pollution from Ships. International Maritime Organization; 2017.
 19. Intership Navigation. Circular Tech#07-Procedures of Implementing MARPOL An. VI. Circular Letters: 2014.
 20. Kaidis C, Uzunoglu B, Amoiralis F. Wind turbine reliability estimation for different assemblies and failure severity categories. *IET Renewable Power Generation* 2015; 9(8): 892–899, <https://doi.org/10.1049/iet-rpg.2015.0020>.
 21. Kamiński W, Krause P, Gumiński D, Rajewski P. The quality of marine fuels and the safety of navigation: case studies. *Scientific Journals of the Maritime University of Szczecin* 2016; 48(120): 15–21, <https://doi.org/10.17402/170>.
 22. Kołodziejcki M. Failure finding tasks in Reliability Centred Maintenance. *Scientific Journals Maritime University of Szczecin* 2011; 28(100): 53–59.
 23. Krystosik-Gromadzińska A. Ship exploitation and rules connected with sulphur limits restrictions. *Scientific Journals Maritime University of Szczecin* 2011; 28(100): 73–77.
 24. Li F, Dang K. Ship's Emission Standards on Fuel Changeover in ECAs (SECAs). In *3rd International Conference on Electromechanical Control Technology and Transportation - ICECTT*, Chongqing: 2018. 347–350, <https://doi.org/10.5220/0006970503470350>.
 25. MAN Diesel & Turbo. Service Letter SL2014-587 / JAP. 2014. [<https://primeserv.man-es.com/marine-engines-and-systems/service-letter-marine>].
 26. MAN Diesel & Turbo. Service Letter SL2014-593/DOJA - Guidelines for Operation on Fuels with less than 0.1% Sulphur. 2014. [<https://primeserv.man-es.com/marine-engines-and-systems/service-letter-marine>].
 27. MAN Diesel & Turbo. Service Letter SL2018-659/JAP - Cermet-Coated Piston Rings for Operation on Low-Sulphur Fuels. 2018. [<https://primeserv.man-es.com/marine-engines-and-systems/service-letter-marine>].
 28. MAN Diesel & Turbo. Service Letter SL2020-692/KAMO - LDCL cooling system update. 2020. [<https://primeserv.man-es.com/marine-engines-and-systems/service-letter-marine>].
 29. MAN Diesel & Turbo. Waste Heat Recovery System (WHRS) for Reduction of Fuel Consumption, Emissions and EEDI. 2014. [<https://mandieselturbo.com/docs/librariesprovider6/technical-papers/waste-heat-recovery-system.pdf>].
 30. Morais C S, Pimenta R E, Ferreira P L et al. Assessing Diabetes Health Literacy, Knowledge and Empowerment in Northern Portugal. *Advances in Intelligent Systems and Computing* 2015; vol 354: 63–71, https://doi.org/10.1007/978-3-319-16528-8_7.
 31. Olaniyi E O, Virmäe M. The Economic Impact of Environmental Regulations on a Maritime Fuel Production Company. *Research in Economics and Business: Central and Eastern Europe* 2016; 8(2): 58–84, [<http://www.rebcee.eu/index.php/REB/article/viewFile/93/77>].
 32. Rausand M, Høyland A. *System Reliability Theory: Models, Statistical Methods, and Applications*. John Wiley and Sons Ltd: 2009. [<https://www.wiley.com/en-us/System+Reliability+Theory%3A+Models+and+Statistical+Methods-p-9780470317747>].
 33. Sasmito H. E. U B. Analisa Keandalan Sistem Bahan Bakar Motor Induk Pada Km. Leuser. *Kapal* 2008; 5(3): 123–135, <https://doi.org/10.12777/kpl.5.2.123-135>.
 34. Shell Marine. IMO 2020 READY. 2019. [https://www.shell.com/business-customers/marine/imo-2020/_jcr_content/par/relatedtopics.stream/1571229884361/2bd59ebc559181c010ae2a4fbec680190ed1409/imo-2020-comprehensive-guide-v17.pdf].
 35. Spinato F, Tavner P J, Van Bussel G J W, Koutoulakos E. Reliability of wind turbine subassemblies. *IET Renewable Power Generation* 2009; 3(4): 387–401, <https://doi.org/10.1049/iet-rpg.2008.0060>.
 36. Wiratama B Y, Nugroho T F, Busse W. Calculation of Temperature Gradient in Manual Fuel Change-Over Operation. *Applied Mechanics and Materials* 2018; vol. 874: 81–87 <https://doi.org/10.4028/www.scientific.net/amm.874.81>.
 37. Witherby Publishing Group. Website. Shipping Regulations and Guidance/Reference/Accident Reports. 2021. [<http://shippingregs.org/Reference/Accident-Reports>].
 38. Zasadzień M. An analysis of the failure frequency of machines in an enterprise characterised by a changeable production level. *Zeszyty Naukowe / Akademia Morska w Szczecinie* 2013; nr 34(106): 103–107.

Reliability modeling for dependent competing failure processes with phase-type distribution considering changing degradation rate

Indexed by:



Hao Lyu^a, Shuai Wang^a, Xiaowen Zhang^a, Zaiyou Yang^a, Michael Pecht^b

^aNortheastern University, School of Mechanical Engineering and Automation, Shenyang 110819, China

^bUniversity of Maryland, Center for Advanced Life Cycle Engineering, College Park, MD 20742, USA


Highlights

- The degradation rate changes when the number of shocks reaches a specific value.
- The phase-type (PH) distribution is combined with the DCFP.
- The survival function of PH distribution is used to calculate hard failure reliability.
- The phase-type distribution method is applied to calculate the reliability of the MEMS.

Abstract

In this paper, a system reliability model subject to Dependent Competing Failure Processes (DCFP) with phase-type (PH) distribution considering changing degradation rate is proposed. When the sum of continuous degradation and sudden degradation exceeds the soft failure threshold, soft failure occurs. The interarrival time between two successive shocks and total number of shocks before hard failure occurring follow the continuous PH distribution and discrete PH distribution, respectively. The hard failure reliability is calculated using the PH distribution survival function. Due to the shock on soft failure process, the degradation rate of soft failure will increase. When the number of shocks reaches a specific value, degradation rate changes. The hard failure is calculated by the extreme shock model, cumulative shock model, and run shock model, respectively. The closed-form reliability function is derived combining with the hard and soft failure reliability model. Finally, a Micro-Electro-Mechanical System (MEMS) demonstrates the effectiveness of the proposed model.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

dependent competing failure processes; phase-type distribution; changing degradation rate; reliability modeling; survival function..

Notation

| | |
|-----------|---|
| $X(t)$ | Continuous degradation at time t |
| $S(t)$ | Cumulative degradation due to random shocks at time t |
| $X_S(t)$ | Total degradation at time t |
| $N(t)$ | Number of random shocks arrived by time t |
| λ | Intensity of random shocks |
| φ | Initial degradation |
| β_1 | Initial degradation rate |
| β_2 | Changed degradation rate when the number of shocks reaches a specific value |
| H | Soft failure threshold |
| D_1 | Hard failure threshold under extreme shock model |
| D_2 | Hard failure threshold under cumulative shock model |

| | |
|----------|--|
| W_L | Critical level on shock magnitude under run shock model |
| W_U | Hard failure threshold under run shock model |
| W_i | The magnitude of the i th shock |
| $F_W(w)$ | Cumulative distribution function (cdf) of W_i |
| Y_i | Degradation damage caused by the i th shock |
| T_j | Arrival time of the j th shock ($T_j \sim \text{Ga}(j, \lambda)$) |
| J | The required number of shocks' occurrences when the soft failure degradation rate changes |
| k | The required number of consecutive shocks that exceed the critical level W_L under run shock model |
| N | The number of transfers before the Markov chain enters the absorption state |
| m | The maximum number of shocks that the system can support |

1. Introduction

Many systems will fail due to various failure modes caused by degradation and random external shocks (such as wear, corrosion, fatigue, fracture, and shock loads) during operation [1]. Some systems may suffer multiple failure processes, and any failure processes will

cause the system to fail. In this paper, we consider two failure processes: soft failure process and hard failure process. Soft failure means that the performance of the system gradually decreases over time. The system will fail when the degradation performance exceeds a certain critical threshold. Common soft failure includes wear, corrosion, and so forth. Hard failure refers to the phenomenon that the system breaks

E-mail addresses: H. Lyu - lvhao@me.neu.edu.cn, S. Wang - 2604150051@qq.com, X. Zhang - 15804042618@163.com, Z. Yang - 1171369333@qq.com, M. Pecht - pecht@umd.edu

down suddenly in the normal working process (e.g., fracture). These two failure processes compete because any failure will cause the system to fail [15]. Besides, because the shock acts on the soft and hard failure processes simultaneously, the soft and hard failure processes are dependent. It is challenging to predict system reliability when the soft and hard failure processes are dependent [25].

Most of researchers are devoted to the reliability prediction of systems that experience degradation or random shocks in the available literature. When there is not enough failure data, the degradation modeling method can indirectly provide the failure information of the system [25]. There are two main types of degradation models: the stochastic process model, such as the Wiener process, Gamma process, and inverse Gaussian process; and the other is the general path model [30]. Ni [21] developed degradation model for a two-stage degradation system subject to shocks, where degradation damage is caused by shocks and follows the Gamma distribution. The general path model is first introduced into the degradation literature by Lu and Meeker [16]. Because it is easy to use and the theory has been well established, the general path model has been used in many DCFP models to describe the degradation process [1, 6, 23, 25]. In our study, in order to implement the idea that the degradation rate changes when the number of shocks reaches a specific value, we use the general path model as the degradation process. Because the degradation rate in the general path model can be changed, this characteristic is exactly consistent with our idea. At the same time, the random shock model has been extensively studied. Various shock models are introduced into the hard failure reliability calculation. Shock models can be divided into the following categories: extreme shock model [29], cumulative shock model [20], run shock model [18], m shock model [11], delta shock model [13], and mixed shock model [26]. In this paper, hard failure is calculated under three different shock patterns: the extreme shock model, the cumulative shock model, and the run shock model.

In the available literature, most of the literature has been devoted to the reliability modeling subject to DCFP. Peng [23] developed reliability modeling for complex systems subject to multiple dependent competing failure processes, where two correlated failure processes are considered. Soft failure is caused jointly by continuous degradation and additional abrupt degradation damage due to a shock process and hard failure caused by abrupt stress from the same shock process. Guo [7] presented a joint copula reliability model for systems experiencing two degradation processes and random shocks, where the dependence between the two degradation processes is considered by copula function. Keedy [12] built a probabilistic reliability model for stents experiencing dependent competing risk processes. Crack propagation is regarded as a degradation process, and a single overload under external shocks is considered a hard failure process. Besides, shocks will accelerate the propagation of cracks, thus forming a dependent competing failure process. Huynh [10] proposed a Degradation-Threshold-Shock model with dependent competing failure modes, where the shock arrival rate follows the nonhomogeneous Poisson process, and the Poisson intensity depends on the degradation level of the system. Jiang [11] established reliability models for systems subject to multiple s-dependent competing failure processes. When the shock meets a particular random shock pattern, the hard failure threshold reduces to a lower level. Rafiee [25] investigated reliability models for a system subject to DCFP of degradation and random shocks with a changing degradation rate according to particular random shock patterns. Lin [14] and Hao [8] studied the general dependences between the degradation and two types of random shocks (extreme shocks and cumulative shocks). Fan [6] established a new reliability model for DCFP, where the intensity function of non-homogeneous Poisson process depends on the degradation processes. Rafiee [26] investigated reliability modeling for systems subject to DCFP considering the impact of a new generalized mixed shock model. When the generalized mixed shock model is satisfied, the degradation rate and the hard failure threshold can simultaneously shift. Zhang [31] proposed a new reliability model for systems with multi-

ple components subject to multiple natural degradations and random shocks, where the degradation rate will accelerate due to shocks. Che [4] studied a novel reliability model for load-sharing k-out-of-n systems, where the dependent workload and shock effects are considered. An [1] considered that systems with high reliability and long life could resist small shocks, and divided shocks into safety shocks, damage shocks, and fatal shocks, and carried out reliability modeling for multiple degradation and shock processes. Lyu [17] applied the reliability model of DCFP to the Turbine and Worm System. Pourhasan [24] put forward a simulation approach about analytic reliability assessment for complicated systems, which embeds the stochastic degradation process and random shocks. In most of the above literature, the interference model is utilized to calculate the hard failure reliability; that is, the system is reliable when the shock magnitude and shock times are less than a certain threshold or the interarrival shock time exceeds a certain threshold. In our research, the phase-type distribution method is employed to calculate the hard failure reliability. The interarrival time between two successive shocks is assumed to be continuous phase-type distribution, and the phase-type distribution survival function is used to calculate the reliability.

The phase-type distribution is suitable for modeling the interarrival time between two successive shocks. There are many advantages about the phase-type distribution method. First, the simplicity of mathematics is one of the advantages of the phase-type distribution method. We can express the distribution and moment in the form of matrix, and it is easy to calculate the results we need [22]. Second, When multiple shock sources act on a system, especially complex shocks such as run shocks, it is difficult to obtain a closed reliability expression with traditional hard failure reliability calculation methods, but the phase-type distribution method is easy to calculate the hard failure reliability. Besides, the closure properties of phase-type distributions under some operations are helpful in the reliability context [2]. In the literature [3, 19, 27], the phase-type distribution is applied to analyze the reliability of shock models. In the literature [3, 19], the interarrival time between shocks is assumed to be continuous phase-type distribution. Shocks may lead to system failure, and the system may fail due to wear. Its wear lifetime follows the continuous phase-type distribution. The interval between shocks and the wear life depend on the number of cumulative shocks. When the shocks are extreme shocks, cumulative shocks, and run shocks, the survival function of the system is obtained. Segovia [27] displayed an analytical expression of the survival function of a multi-state system that suffered shocks by using phase-type distribution. Zhao [32] proposed a multi-state shock model, where the Markov chain was constructed by the number of shocks of different types of shocks. When the interarrival time between shocks follows the common continuous phase-type distribution, the survival function and mean residual lifetime of the multi-state system were derived. Eryilmaz [5] developed a new mixed shock model, which combined the extreme shock model and the run shock model. The survival function of the system was studied when the interarrival time and the shock magnitude are independent and dependent using the property of phase-type distribution. The above literature used the phase-type distribution calculation method when calculating the failure reliability caused by shocks. Literature [3, 19, 27] assumed that the wear life follows a continuous phase-type distribution. When the failure is caused by wear, the survival function is used to describe the reliability of the system.

In the existing literature, the phase-type distribution calculation method has not been combined with DCFP. We combine the phase-type distribution calculation method with DCFP. As far as the author's knowledge, this is the first time for the research in combining phase-type distribution with DCFP. In this paper, the general path model is utilized for the soft failure. The degeneration path is assumed to be a linear path. The degradation rate changes when the number of shocks reaches a specific value. The soft failure reliability is calculated by the total degradation-threshold interference model. The phase-type distribution method is applied to calculate the hard failure. It is as-

sumed that the interarrival time between shocks follows the common phase-type distribution, the total number of shocks before the hard failure occurring follows the discrete phase-type distribution, and the survival function is employed to calculate the hard failure reliability. The total reliability of the system is derived considering the hard and soft failures by the number of shocks.

The rest of this article is organized as follows. In Section 2, the soft failure process and the hard failure process of the system are described, along with the dependent competing failure relationship between those processes. In Section 3, the reliability models of the system, including the soft failure model, the hard failure model (the extreme shock model, cumulative shock model, and run shock model), and the model of DCFP, are established. In Section 4, a numerical example is developed to demonstrate the implementation and effectiveness of the proposed model. In Section 5, the calculation results are summarized.

2. System description and preliminaries

As shown in Figure 1, the failure of a system is caused by two dependent competing failure processes: the soft failure process and the hard failure process. The total degradation of the soft failure process consists of continuous degradation and sudden degradation caused by shocks. When the total degradation exceeds the soft failure threshold H , soft failure occurs in the system. At the same time, hard failure will occur when the shock magnitude exceeds the hard failure threshold D . Whichever failure processes occurs first will cause the system to fail. The shock process acts on the soft and hard failure process simultaneously, so system failure results from dependent competing failures in the soft and hard failure process. In this paper, three shock models are applied for the hard failure process: (1) Extreme shock model, when the shock magnitude exceeds the hard failure threshold, the system will have a hard failure. (2) Cumulative shock model, when the cumulative magnitude of shocks exceeds the hard failure threshold, hard failure occurs in the system. (3) Run shock model, when the magnitude of k consecutive shocks exceeds the critical threshold, hard failure occurs. Besides, when the number of shocks reaches a certain value, the degradation rate of soft failure changes.

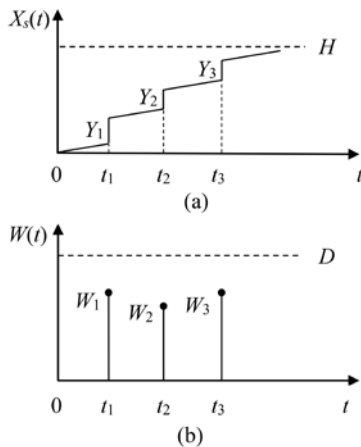


Fig. 1. Two dependent competing failure processes: (a) soft failure, (b) hard failure

Phase-type distributions and property:

Consider a finite discrete-time Markov chain in the state space $\{1, 2, \dots, m, m+1\}$, where $1, 2, \dots, m$ are the transient states, and $m+1$ is the absorbing state. The number of transitions before the Markov chain enters the absorbing state is defined as a discrete phase-type distribution. The probability mass function of discrete phase-type distributed random variable N is [22]:

$$P\{N=n\} = \mathbf{a}\mathbf{Q}^{n-1}\mathbf{u}', \quad n=1,2,\dots \quad (1)$$

where, for $n \in \mathbb{N}$, $\mathbf{Q}=(q_{ij})_{m \times m}$ is the transition probability matrix between m transient states, and $\mathbf{u}'=(\mathbf{I}-\mathbf{Q})\mathbf{e}'$ is the transition probability vector from the transient state to the absorbing state, \mathbf{I} is the identity matrix. The matrix \mathbf{Q} must satisfy the condition that $\mathbf{I}-\mathbf{Q}$ is non-singular. We use $N \sim PH_d(\mathbf{a}, \mathbf{Q})$ to indicate that the random variable N follows the discrete phase-type distribution.

Assuming a finite-state Markov process starts the transition from transient state i with probability a_i . The time distribution of the Markov process entering the absorbing state is defined as continuous phase-type distribution. The cumulative distribution function of continuous phase-type distributed random variable X is [22]:

$$P(X \leq x) = 1 - \alpha \exp(\mathbf{A}x)\mathbf{e}' \quad (2)$$

The survival function of X is given by:

$$P(X > x) = \alpha \exp(\mathbf{A}x)\mathbf{e}' \quad (3)$$

where, \mathbf{A} is an $m \times m$ matrix, whose diagonal elements are negative, and non-diagonal elements are non-negative, and $\mathbf{e}=(1, \dots, 1)_{1 \times m}$. All elements of the row vector $\alpha=(a_1, \dots, a_m)$ are non-negative. Exponential, Erlang, generalized Erlang, and Coxian distributions are commonly-used continuous phase-type distributions [9]. We use $X \sim PH_c(\alpha, \mathbf{A})$ to indicate that the random variable X follows the continuous phase-type distribution of order m with a PH-generator \mathbf{A} and substochastic vector α .

Proposition [22]: Assume that X_1, X_2, \dots are independent and $X_i \sim PH_c(\alpha, \mathbf{A})$, $i=1,2,\dots$ and independently $N \sim PH_d(\mathbf{a}, \mathbf{Q})$. If α and \mathbf{a} are stochastic vectors, i.e., $\alpha\mathbf{e}'=1, \mathbf{a}\mathbf{e}'=1$, then $\sum_{i=1}^N X_i \sim PH_c(\alpha \otimes \mathbf{a}, \mathbf{A} \otimes \mathbf{I} + (\mathbf{a}^0 \alpha) \otimes \mathbf{Q})$, $\mathbf{a}^0 = -\mathbf{A}\mathbf{e}'$.

where \otimes is the Kronecker product.

3. Reliability analysis of DCFP considering time phase-type distribution

In this section, the reliability analysis of the system experiencing the degradation process and the shock process is carried out. First, the soft failure model is developed—the degradation rate changes when the number of shocks reaches a specific value. Then the phase-type distribution is utilized to model the hard failure process (including extreme shock, cumulative shock, and run shock). Finally, the total reliability is calculated.

3.1. Soft failure model under degradation and random shocks

The total degradation includes continuous degradation and abrupt degradation caused by random shocks. The continuous degradation path is assumed to be a linear path $X(t)=\varphi+\beta t$. φ is the initial degradation, β_1 is the degradation rate of the first stage, β_2 is the degradation rate of the second stage. Assume that the initial degradation φ , the degradation rate β_1 and β_2 all follow the normal distribution, that is, $\beta_1 \sim N(\mu_{\beta_1}, \sigma_{\beta_1}^2)$, $\beta_2 \sim N(\mu_{\beta_2}, \sigma_{\beta_2}^2)$. The degradation rate changes when the j th shock arrives. Then the continuous degradation $X(t)$ can be expressed as:

$$X(t) = \begin{cases} \varphi + \beta_1 t, & j > N(t) \\ \varphi + \beta_1 T_j + \beta_2 (t - T_j), & j \leq N(t) \end{cases} \quad (4)$$

where, T_j is the time of arrival of the j th shock, and $N(t)$ is the number of random shocks arrived by time t .

Random shocks will cause abrupt degradation damage to the degradation process, thereby accelerating the degradation process. Assuming that the magnitude of the random shock W_i is independent and identically normally distributed, namely $W_i(t_i) \sim N(\mu_W, \sigma_W^2)$, t_i is the arrival time of the i th shock. The cumulative distribution function of the shock magnitude is $F_W(x)$. The arrival times of random shocks follow a homogeneous Poisson process with intensity λ , then:

$$P\{N(t) = i\} = \frac{(\lambda t)^i}{i!} e^{-\lambda t}, \quad i = 0, 1, 2, \dots \quad (5)$$

When the number of random shock arrivals follows the Poisson process with intensity λ , for a certain j , the arrival time T_j of the j th shock follows the Gamma distribution with shape parameter j and scale parameter λ , that is, $T_j \sim \text{Ga}(j, \lambda)$. The probability density function is:

$$f_{T_j}(t_j; j) = \frac{\lambda^j}{(j-1)!} t_j^{j-1} e^{-\lambda t_j} \quad (6)$$

Let $Y_i (i=1, 2, \dots, \infty)$ be the abrupt degeneration increment caused by the i th random shock, that is, the damage caused by the random shock to the degradation process. Then the total degradation $S(t)$ caused by random shocks is:

$$S(t) = \begin{cases} \sum_{i=1}^{N(t)} Y_i, & N(t) > 0 \\ 0, & N(t) = 0 \end{cases} \quad (7)$$

Then the total degradation of soft failure $X_s(t)$ can be expressed as:

$$X_s(t) = X(t) + S(t) \quad (8)$$

To keep the system in normal working condition, the total degradation of the system $X_s(t)$ should be less than the soft failure critical threshold H . The reliability of the soft failure is:

$$\begin{aligned} R_s(t) &= P(X_s(t) < H) = P(X(t) + S(t) < H) = P\left(X(t) + \sum_{i=1}^{N(t)} Y_i < H\right) \\ &= P(X(t) < H | N(t) = 0) \cdot P(N(t) = 0) + \sum_{i=1}^{\infty} P\left(X(t) + \sum_{i=1}^{N(t)} Y_i < H | N(t) = i\right) \cdot P(N(t) = i) \\ &= P(\varphi + \beta_1 t < H) \cdot P(N(t) = 0) + \sum_{i=1}^j P\left(\varphi + \beta_1 t + \sum_{i=1}^{N(t)} Y_i < H\right) \cdot P(N(t) = i) \\ &+ \sum_{i=j+1}^{\infty} P\left(\varphi + \beta_1 T_j + \beta_2(t - T_j) + \sum_{i=1}^{N(t)} Y_i < H\right) \cdot P(N(t) = i) \\ &= \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t)}{\sqrt{\sigma_{\beta_1}^2 t^2}}\right) \cdot \exp(-\lambda t) + \sum_{i=1}^j \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t + i\mu_Y)}{\sqrt{\sigma_{\beta_1}^2 t^2 + i\sigma_Y^2}}\right) \cdot \frac{\exp(-\lambda t)(\lambda t)^i}{i!} \\ &+ \sum_{i=j+1}^{\infty} \int_0^t \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t_j + \mu_{\beta_2}(t - t_j) + i\mu_Y)}{\sqrt{\sigma_{\beta_1}^2 t_j^2 + \sigma_{\beta_2}^2(t - t_j)^2 + i\sigma_Y^2}}\right) \cdot \frac{\lambda^j}{(j-1)!} t_j^{j-1} e^{-\lambda t_j} dt_j \cdot \frac{\exp(-\lambda t)(\lambda t)^i}{i!} \end{aligned} \quad (9)$$

3.2. Hard failure model under extreme shock

The extreme shock model is shown in Figure 2. It can be seen from Figure 2 (a) that when the number of shocks reaches a specific value (the schematic diagram is 3), the soft failure degradation rate increases from β_1 to β_2 . As shown in Figure 2(b), the fourth shock is a fatal shock, so the system life is $T = X_1 + X_2 + X_3 + X_4$.

Let X_i denote the interarrival time between the i th shock and the $i-1$ th shock, $i \geq 1$. Suppose the arrival rate of the shock follows a Pois-

son distribution with parameter λ . In that case, the interarrival time X_i follows the exponential distribution, which can be expressed as phase-type distribution:

$$X_i \sim PH_c(\mathbf{a}, \mathbf{A}) = PH_c(1, -\lambda) \quad (10)$$

Let p_1 be the probability of a fatal shock, and $1-p_1$ be the probability of a non-fatal shock.

$$p_1 = P(W_i > D_1) = 1 - F_W(D_1) = 1 - \Phi\left(\frac{D - \mu_W}{\sigma_W}\right) \quad (11)$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.

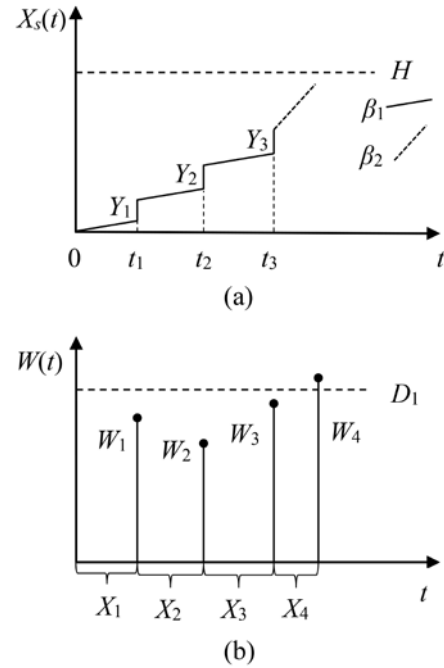


Fig. 2. Extreme shock model

Let N be the number of transfers before the Markov chain enters the absorption state, that is, the number of shocks before hard failure occurring, which follows the discrete phase-type distribution, namely $N \sim PH_d(\mathbf{a}, \mathbf{Q})$

$$\mathbf{a} = (1 \ 0 \ 0 \ 0 \ \dots \ 0)_{1 \times (m+1)}, \quad \mathbf{Q} = \begin{pmatrix} 0 & 1-p_1 & 0 & \dots & 0 \\ 0 & 0 & 1-p_1 & \dots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1-p_1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}_{(m+1) \times (m+1)} \quad (12)$$

where, m is the maximum number of shocks that the system can support.

Let T be the life of the hard failure of the system, then according to the phase-type distribution properties, we have:

$$T = \sum_{i=1}^N X_i \sim PH_c(\mathbf{g}, \mathbf{G}) = PH_c\left(\pm \otimes \mathbf{a}, \mathbf{A} \otimes \mathbf{I} + (\mathbf{a}^0 \alpha) \otimes \mathbf{Q}\right), \quad \mathbf{a}^0 = -\mathbf{A} \mathbf{e}'$$

$$\mathbf{g} = (1 \ 0 \ 0 \ 0 \ \dots \ 0)_{1 \times (m+1)}, \quad \mathbf{G} = \begin{pmatrix} -\lambda & \lambda(1-p_1) & 0 & \dots & 0 \\ 0 & -\lambda & \lambda(1-p_1) & \dots & 0 \\ 0 & 0 & -\lambda & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \lambda(1-p_1) \\ 0 & 0 & 0 & \dots & -\lambda \end{pmatrix}_{(m+1) \times (m+1)} \quad (13)$$

According to the phase-type distribution survival function, we have:

$$P(T > t) = \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \quad (14)$$

Because the soft failure reliability formula is derived by the number of shocks as the conditional probability, in order to unify the reliability expression of soft and hard failures, the hard failure reliability formula also uses the number of shocks as the conditional probability. Therefore, the hard failure reliability can be expressed as:

$$\begin{aligned} R_H(t) &= P(T > t) = \sum_{i=0}^{\infty} (T > t | N(t) = i) \cdot P(N(t) = i) \\ &= \sum_{i=0}^{\infty} P(T > t) \cdot \frac{\exp(-\lambda t) (\lambda t)^i}{i!} = \sum_{i=0}^{\infty} \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \frac{\exp(-\lambda t) (\lambda t)^i}{i!} \end{aligned} \quad (15)$$

3.3. Hard failure model under cumulative shock

The cumulative shock model is shown in Figure 3. It can be seen from Figure 3 (a) that when the number of shocks reaches a specific value (the schematic diagram is 3), the soft failure degradation rate increases from β_1 to β_2 . As shown in Figure 3(b), the fourth cumulative shock exceeds the hard failure threshold, the system fails, so the system life is $T = X_1 + X_2 + X_3 + X_4$.

Let p_i be the probability that the i th cumulative shock is a fatal shock, then:

$$p_i = 1 - \Phi\left(\frac{D_2 - i\mu_W}{\sqrt{i\sigma_W^2}}\right) \quad (16)$$

Let N be the number of transfers before the Markov chain enters the absorption state, which follows the discrete phase-type distribution, namely $N \sim PH_d(\mathbf{a}, \mathbf{Q})$

$$\mathbf{a} = (1 \ 0 \ 0 \ 0 \ \dots \ 0)_{1 \times (m+1)}, \quad \mathbf{Q} = \begin{pmatrix} 0 & 1-p_1 & 0 & \dots & 0 \\ 0 & 0 & 1-p_2 & \dots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1-p_m \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}_{(m+1) \times (m+1)} \quad (17)$$

where, m is the maximum number of shocks that the system can support.

Let T be the life of the hard failure of the system, then according to the phase-type distribution properties, we have:

$$T = \sum_{i=1}^N X_i \sim PH_c(\mathbf{g}, \mathbf{G}) = PH_c\left(\pm \otimes \mathbf{a}, \mathbf{A} \otimes \mathbf{I} + (\mathbf{a}^0 \alpha) \otimes \mathbf{Q}\right), \quad \mathbf{a}^0 = -\mathbf{A} \mathbf{e}'$$

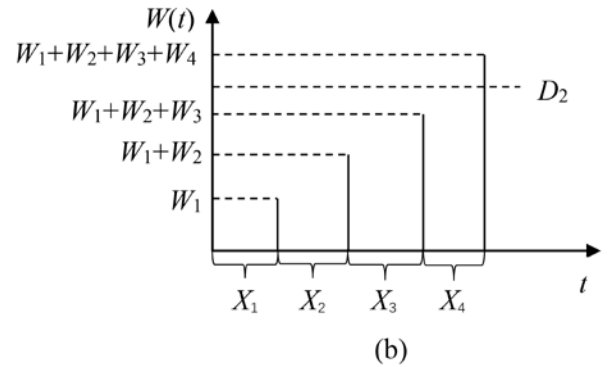
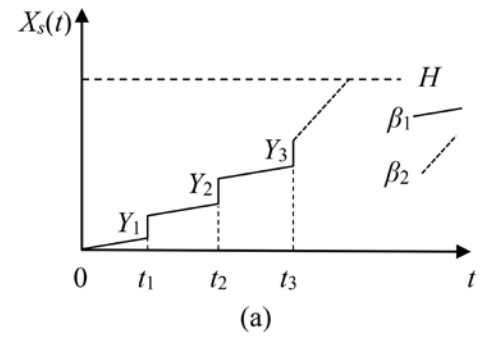


Fig. 3. Cumulative shock model

$$\mathbf{g} = (1 \ 0 \ 0 \ 0 \ \dots \ 0)_{1 \times (m+1)}, \quad \mathbf{G} = \begin{pmatrix} -\lambda & \lambda(1-p_1) & 0 & \dots & 0 \\ 0 & -\lambda & \lambda(1-p_2) & \dots & 0 \\ 0 & 0 & -\lambda & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \lambda(1-p_m) \\ 0 & 0 & 0 & \dots & -\lambda \end{pmatrix}_{(m+1) \times (m+1)} \quad (18)$$

According to the phase-type distribution survival function, we have $P(T > t) = \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}'$

Therefore, the hard failure reliability can be expressed as:

$$\begin{aligned} R_H(t) &= P(T > t) = \sum_{i=0}^{\infty} (T > t | N(t) = i) \cdot P(N(t) = i) \\ &= \sum_{i=0}^{\infty} P(T > t) \cdot \frac{\exp(-\lambda t) (\lambda t)^i}{i!} = \sum_{i=0}^{\infty} \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \frac{\exp(-\lambda t) (\lambda t)^i}{i!} \end{aligned} \quad (19)$$

3.4. Hard failure model under run shock

The run shock model is shown in Figure 4. $k=2$ means that when the magnitude of two consecutive shocks exceeds the critical threshold W_L , hard failure occurs. It can be seen from Figure 4 (a) that when the number of shocks reaches a specific value (the schematic diagram is 3), the soft failure degradation rate increases from β_1 to β_2 . As shown in Figure 4(b), when the fourth shock arrives, the condition of system failure caused by run shock is met, so the system life is $T = X_1 + X_2 + X_3 + X_4$.

Let p represent the probability that the shock exceeds the critical level of run shock model under the condition that the shock is not fatal, then:

$$p = P(W_i > W_L | W_i < W_U) = \frac{P(W_L < W_i < W_U)}{P(W_i < W_U)} = \frac{F_W(W_U) - F_W(W_L)}{F_W(W_U)} \quad (20)$$

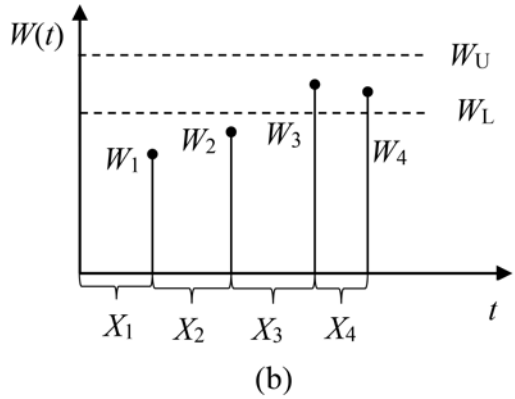
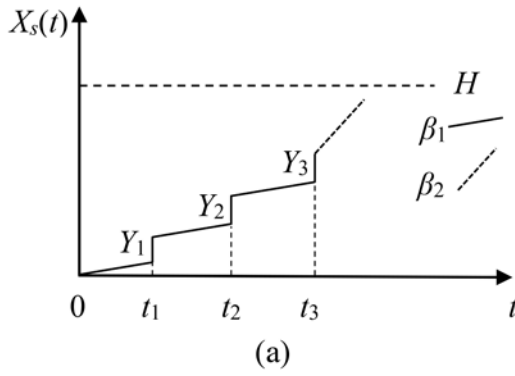


Fig. 4. Run shock model($k=2$)

Let N be the number of transfers before the Markov chain enters the absorption state, which follows the discrete phase-type distribution, namely $N \sim PH_d(\mathbf{a}, \mathbf{Q})$.

$$\mathbf{a} = (1 \ 0 \ 0 \ 0 \ \cdots \ 0)_{1 \times k}, \quad \mathbf{Q} = \begin{pmatrix} 1-p & p & 0 & \cdots & 0 \\ 1-p & 0 & p & \cdots & 0 \\ 1-p & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & p \\ 1-p & 0 & 0 & \cdots & 0 \end{pmatrix}_{k \times k} \quad (21)$$

where, k is the required number of consecutive shocks that exceed the critical level W_L under run shock model.

Let T be the life of the hard failure of the system, then according to the phase-type distribution properties, we have:

$$T = \sum_{i=1}^N X_i \sim PH_c(\mathbf{g}, \mathbf{G}) = PH_c(\mathbf{a} \otimes \mathbf{a}, \mathbf{A} \otimes \mathbf{I} + (\mathbf{a}^0 \mathbf{a}) \otimes \mathbf{Q}), \quad \mathbf{a}^0 = -\mathbf{A} \mathbf{e}'$$

$$\mathbf{g} = (1 \ 0 \ 0 \ 0 \ \cdots \ 0)_{1 \times k}, \quad \mathbf{G} = \begin{pmatrix} -\lambda p & \lambda p & 0 & \cdots & 0 \\ \lambda(1-p) & -\lambda & \lambda p & \cdots & 0 \\ \lambda(1-p) & 0 & -\lambda & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \lambda p \\ \lambda(1-p) & 0 & 0 & \cdots & -\lambda \end{pmatrix}_{k \times k} \quad (22)$$

According to the phase-type distribution survival function, we have $P(T > t) = \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}'$

Therefore, the hard failure reliability can be expressed as:

$$\begin{aligned} R_H(t) &= P(T > t) = \sum_{i=0}^{\infty} (T > t | N(t) = i) \cdot P(N(t) = i) \\ &= \sum_{i=0}^{\infty} P(T > t) \cdot \frac{\exp(-\lambda t)(\lambda t)^i}{i!} = \sum_{i=0}^{\infty} \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \frac{\exp(-\lambda t)(\lambda t)^i}{i!} \end{aligned} \quad (23)$$

3.5. System reliability analysis

The system experiences both soft and hard failure processes at the same time. If the system is not to fail, neither soft nor hard failures can occur. According to Section 3.1 to 3.4, we have obtained the system's soft and hard failure reliability expressions. Therefore the total reliability is:

$$\begin{aligned} R(t) &= P(X_s(t) < H, T > t) = P(X_s(t) < H, T > t | N(t) = 0) \cdot P(N(t) = 0) \\ &+ \sum_{i=1}^{\infty} P(X_s(t) < H, T > t | N(t) = i) \cdot P(N(t) = i) \\ &= P(X(t) < H, T > t | N(t) = 0) \cdot P(N(t) = 0) + \sum_{i=1}^{\infty} P(X_s(t) < H, T > t | N(t) = i) \cdot P(N(t) = i) \\ &= \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t)}{\sqrt{\sigma_{\beta_1}^2 t^2}}\right) \cdot \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \exp(-\lambda t) + \sum_{i=1}^j P\left(\varphi + \beta_1 t + \sum_{l=1}^{N(t)} Y_l < H\right) \cdot P(T > t) \cdot P(N(t) = i) \\ &+ \sum_{i=j+1}^{\infty} P\left(\varphi + \beta_1 T_j + \beta_2(t - T_j) + \sum_{l=1}^{N(t)} Y_l < H\right) \cdot P(T > t) \cdot P(N(t) = i) \\ &= \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t)}{\sqrt{\sigma_{\beta_1}^2 t^2}}\right) \cdot \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \exp(-\lambda t) + \sum_{i=1}^j \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t + i \mu_Y)}{\sqrt{\sigma_{\beta_1}^2 t^2 + i \sigma_Y^2}}\right) \cdot \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \frac{\exp(-\lambda t)(\lambda t)^i}{i!} \\ &+ \sum_{i=j+1}^{\infty} \int_0^t \Phi\left(\frac{H - (\varphi + \mu_{\beta_1} t_j + \mu_{\beta_2}(t - t_j) + i \mu_Y)}{\sqrt{\sigma_{\beta_1}^2 t_j^2 + \sigma_{\beta_2}^2(t - t_j)^2 + i \sigma_Y^2}}\right) \frac{\lambda^j}{(j-1)!} t_j^{j-1} e^{-\lambda t_j} dt_j \cdot \mathbf{g} \exp(\mathbf{G}t) \mathbf{e}' \cdot \frac{\exp(-\lambda t)(\lambda t)^i}{i!} \end{aligned} \quad (24)$$

4. Numerical examples

In this section, a micro-engine is studied as a realistic example to illustrate the proposed model's effectiveness in this paper. The micro-engine includes comb-drive actuators and rotating gear, which are mechanically connected. After the voltage is applied, the comb-drive linear displacement is transformed into the circular motion of the gear through the pin joint. According to the experimental research conducted by Sandia National Laboratory, the wear of the friction surface between the gear and the cylindrical pin is the primary failure mode of the micro-engine, and the increase in wear eventually causes the cylindrical pin to break. The micro-engine is not only subjected to wear but also to random shocks. Tanner *et al.* [28] conducted a reliability analysis on the micro-engine in the shock environment. Random shocks will cause wear debris and accelerate the wear of the friction surface. Besides, under the impact of the shock, the spring may be misaligned, and a shock with sufficient magnitude may cause the spring to break. Because the shock will accelerate the degradation process, we assume that the degradation rate increases after the number of shocks reaches a specific value. The parameters used in reliability analysis are shown in Table 1.

The total reliability curves, soft failure reliability curves, and hard failure reliability curves under the extreme shock model, cumulative shock model, and run shock model are shown in Fig. 5. Besides, the sensitivity curves of (D_1, D_2, W_L) , Poisson intensity λ , and soft failure degradation rate β_2 under three shock models are demonstrated in Fig. 6 – 8.

It can be seen from the soft failure reliability curve and the total reliability curve in Fig. 5 (a) that when t is around 0.8×10^5 , the decline rate of the soft failure reliability curve and the total reliability curve becomes faster, which is because the number of shock arrivals reaches a certain threshold at this time. The soft failure degradation rate in-

Table 1. Parameter values of the reliability model

| Parameters | Values | Sources |
|--------------------|--|----------------------|
| H | $0.00125 \mu\text{m}^3$ | (Tanner&Dugger,2003) |
| D_1 | 1.5 GPa | (Rafiee, 2014) |
| D_2 | 5.0 GPa | (Hao, 2017) |
| W_U | 1.8 GPa | Assumption |
| W_L | 1.5 GPa | (Rafiee, 2014) |
| φ | 0 | (Tanner&Dugger,2003) |
| $\mu_{\beta 1}$ | $8.4823 \times 10^{-9} \mu\text{m}^3$ | (Tanner&Dugger,2003) |
| $\sigma_{\beta 1}$ | $6.0016 \times 10^{-10} \mu\text{m}^3$ | (Tanner&Dugger,2003) |
| $\mu_{\beta 2}$ | $10.4823 \times 10^{-9} \mu\text{m}^3$ | (Rafiee, 2014) |
| $\sigma_{\beta 2}$ | $6.0016 \times 10^{-10} \mu\text{m}^3$ | (Tanner&Dugger,2003) |
| μ_W | 1.2 GPa | (Rafiee, 2014) |
| σ_W | 0.2 GPa | (Rafiee, 2014) |
| μ_Y | $1.0 \times 10^{-4} \mu\text{m}^3$ | (Rafiee, 2014) |
| σ_Y | $2 \times 10^{-5} \mu\text{m}^3$ | (Rafiee, 2014) |
| λ | 5×10^{-5} / revolutions | (Rafiee, 2014) |
| j | 3 | Assumption |
| k | 2 | Assumption |

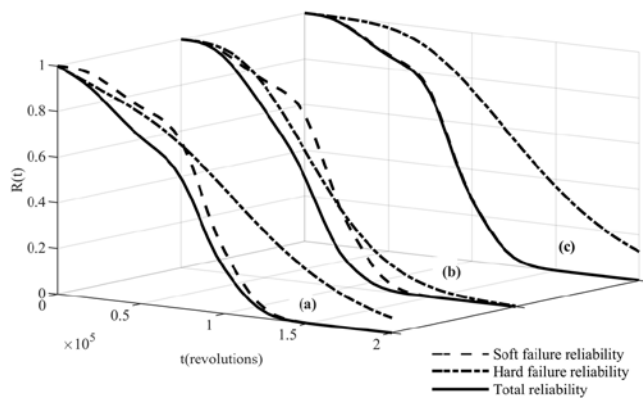


Fig. 5. Curves of soft failure reliability, hard failure reliability, and total reliability: (a) extreme shock model, (b) cumulative shock model, (c) run shock model

creases, resulting in a faster decline in soft failure reliability and total reliability. It can be seen from Fig. 6 (a) that when the hard failure threshold D_1 is increased from 1.3 to 1.6, the reliability curve shifts to the right. We have increased the hard failure threshold, and the system has better performance, which increases the hard failure reliability, and the total reliability becomes greater. It can be seen from Fig. 7 (a) that when the Poisson intensity λ increases, the reliability curve shifts to the left. We have increased the frequency of shock arrivals, and the system is in a worse working environment, thus reducing the reliability. It can be seen from Fig. 8 (a) that with the increase of β_2 , the reliability curve shifts to the left, which is due to the increase in the rate of soft failure degradation leads to a decrease in the soft failure reliability, thereby reducing the total reliability.

It can be seen from Fig. 6 (b) that when the hard failure threshold D_2 increases from 4.0 to 7.0, the total reliability curve shifts to the right. As D_2 decreases, the inflection point of the total reliability curve becomes less noticeable. It is because when D_2 is a smaller value, the number of shocks required to cause the system to fail is small. The system will fail when the number of shocks has not reached a predetermined value that changes the degradation rate of soft failure, so the

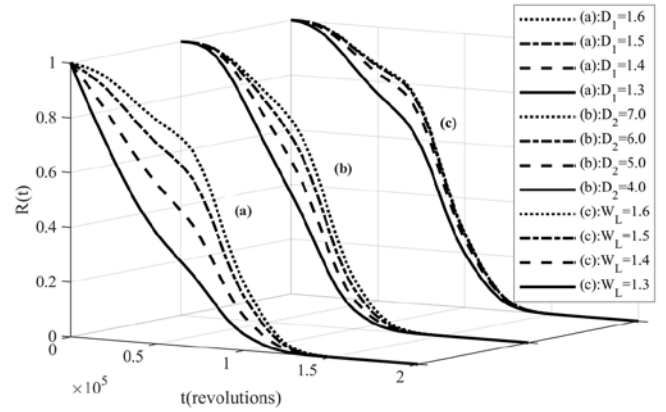


Fig. 6. Sensitivity analysis of $R(t)$ on D_1 , D_2 , W_L : (a) extreme shock model, (b) cumulative shock model, (c) run shock model

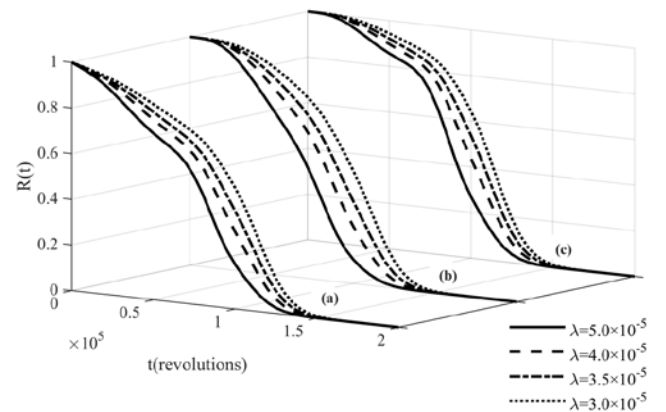


Fig. 7. Sensitivity analysis of $R(t)$ on λ : (a) extreme shock model, (b) cumulative shock model, (c) run shock model

inflection point of the reliability curve is not apparent. It can be seen from Fig. 7 (b) that when the Poisson intensity λ increases, the reliability curve shifts more obviously to the left, which indicates that the reliability of the system is more sensitive to the frequency of shock arrival. So it is necessary to minimize the frequency of shock arrivals to maintain high reliability. It can be seen from Fig. 8 (b) that with the increase of β_2 , the reliability curve shifts to the left. The rise of the soft failure degradation rate leads to a decrease in soft failure reliability, reducing the total reliability.

It can be seen from Fig. 5 (c) that compared with the total reliability under the extreme shock model (see Fig. 5 (a)), the total reliability under the run shock model is higher. It is because when the hard failure threshold under extreme shock model D_1 and the critical level on shock magnitude under run shock model W_L are the same, the run shock model requires that the system fails when two consecutive shocks exceed W_L , while the extreme shock model only needs one shock to exceed D_1 . It can be seen from Fig. 6 (c) that when the critical level on shock magnitude W_L increases from 1.3 to 1.6, the total reliability curves are relatively close, which shows that the reliability of the system is less sensitive to the critical level on shock magnitude W_L . It can be seen from Fig. 7 (c) that when the Poisson intensity λ increases, the reliability curve shifts more obviously to the left, which indicates that the reliability of the system is more sensitive to the frequency of the shock. So it is necessary to minimize the frequency of shock arrivals to maintain high reliability. It can be seen from Fig. 8 (c) that with the increase of β_2 , the reliability curve shifts to the left. It is due to the rise in the soft failure degradation rate, which leads to a decrease in the soft failure reliability and the total reliability.

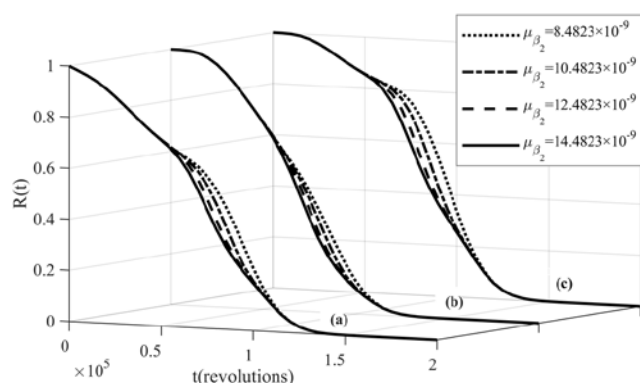


Fig. 8. Sensitivity analysis of $R(t)$ on β_2 : (a) extreme shock model, (b) cumulative shock model, (c) run shock model

5. Conclusions

In this paper, based on the phase-type distribution, we develop a new reliability model for systems subject to DCFP with phase-type distribution considering changing degradation rate. The main innova-

tions of this paper are as follows: first, when the number of shocks reaches a specific value, the soft failure degradation rate changes; second, the phase-type distribution method is utilized to calculate the hard failure reliability—the interarrival time between two successive shocks follows a continuous phase-type distribution, and the survival function of the phase-type distribution is applied to calculate the hard failure reliability; third, the phase-type distribution is combined with the DCFP. Besides, the hard failure shock model adopts the extreme shock model, cumulative shock model, and run shock model, respectively. Finally, the proposed new model is verified by a MEMS numerical example. The effect of model parameters is studied through sensitivity analysis.

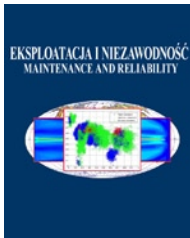
Acknowledgment

This research was funded by the National Natural Science Foundation of China Project (51605083, 12072069); supported by the fundamental research funds for the Central Universities of China Project (N180304022); supported by China Scholarship Council Visiting Scholars Project (201906085037).

References

1. An Z W, Sun D M. Reliability modeling for systems subject to multiple dependent competing failure processes with shock loads above a certain level. *Reliability Engineering & System Safety* 2017; 157: 129–138, <https://doi.org/10.1016/j.res.2016.08.025>.
2. Assaf D, Levikson B. Closure of Phase Type Distributions Under Operations Arising in Reliability Theory. *Annals of Probability* 1982; 10:265–269, <https://doi.org/10.1214/aop/1176993932>.
3. Cazorla D, Pérez-Ocón R, Segovia García M. Survival Probabilities for Shock and Wear Models Governed by Phase-Type Distributions. *Quality Technology & Quantitative Management* 2007; 4: 85–94, <https://doi.org/10.1080/16843703.2007.11673136>.
4. Che H, Zeng S, Guo J. A reliability model for load-sharing k-out-of-n systems subject to soft and hard failures with dependent workload and shock effects. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2020; 22 (2): 253–264, <http://dx.doi.org/10.17531/ein.2020.2.8>.
5. Eryilmaz S, Tekin M. Reliability evaluation of a system under a mixed shock model. *Journal of Computational and Applied Mathematics* 2019; 352: 255–261, <https://doi.org/10.1016/j.cam.2018.12.011>.
6. Fan M F, Zeng Z G, Zio E, Kang R. Modeling dependent competing failure processes with degradation-shock dependence. *Reliability Engineering & System Safety* 2017; 165: 422–430, <https://doi.org/10.1016/j.res.2017.05.004>.
7. Guo C, Wang W, Guo B, Peng R. Maintenance Optimization for Systems With Dependent Competing Risks Using a Copula Function. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2013; 15 (1): 9–17.
8. Hao S H, Yang J, Ma X B, Zhao Y. Reliability modeling for mutually dependent competing failure processes due to degradation and random shocks. *Applied Mathematical Modelling* 2017; 51: 232–249, <https://doi.org/10.1016/j.apm.2017.06.014>.
9. He Q M. *Fundamentals of Matrix-Analytic Methods*. New York: Springer, 2014, <https://doi.org/10.1007/978-1-4614-7330-5>.
10. Huynh K T, Castro I, Barros A, Berenguer C. Modeling age-based maintenance strategies with minimal repairs for systems subject to competing failure modes due to degradation and shocks. *European Journal of Operational Research* 2012; 218: 140–151, <https://doi.org/10.1016/j.ejor.2011.10.025>.
11. Jiang L, Feng Q M, Coit D W. Reliability and Maintenance Modeling for Dependent Competing Failure Processes With Shifting Failure Thresholds. *IEEE Transactions on Reliability* 2012; 61: 932–948, <https://doi.org/10.1109/TR.2012.2221016>.
12. Keedy E, Feng Q M. Reliability Analysis and Customized Preventive Maintenance Policies for Stents With Stochastic Dependent Competing Risk Processes. *IEEE Transactions on Reliability* 2013; 62: 887–897, <https://doi.org/10.1109/TR.2013.2285045>.
13. Li Z, Kong X. Life behavior of δ -shock model. *Statistics & Probability Letters* 2007; 77 (6): 577–587, <https://doi.org/10.1016/j.spl.2006.08.008>.
14. Lin Y H, Li Y F, Zio E. Integrating Random Shocks Into Multi-State Physics Models of Degradation Processes for Component Reliability Assessment. *IEEE Transactions on Reliability* 2014; 64: 154–166, <https://doi.org/10.1109/TR.2014.2354874>.
15. Lin Y H, Li Y F, Zio E. Reliability Assessment of Systems Subject to Dependent Degradation Processes and Random Shocks. *IIE Transactions* 2016; 48: 1072–1085, <https://doi.org/10.1080/0740817X.2016.1190481>.
16. Lu C, Meeker W. Using Degradation Measures to Estimate a Time-to-Failure Distribution. *Technometrics* 1993; 35: 161–174, <https://doi.org/10.2307/1269661>.
17. Lyu H, Zhang X W, Yang Z Y, et al. Reliability Analysis for the Dependent Competing Failure with Wear Model and Its Application to the Turbine and Worm System. *IEEE Access* 2021 (9): 50265–50280, <https://doi.org/10.1109/ACCESS.2021.3062026>.
18. Mallor F, Omey E. Shocks, runs and random sums. *Journal of Applied Probability* 2001; 38: 438–448, <https://doi.org/10.1239/jap/996986754>.
19. Montoro-Cazorla D, Pérez-Ocón R, Segovia M C. Shock and wear models under policy N using phase-type distributions. *Applied Mathematical Modelling* 2009; 33 (1): 543–554, <https://doi.org/10.1016/j.apm.2007.11.017>.
20. Montoro-Cazorla D, Pérez-Ocón R. A reliability system under cumulative shocks governed by a BMAP. *Applied Mathematical Modelling* 2015; 39 (23): 7620–7629, <https://doi.org/10.1016/j.apm.2015.03.066>.

21. NI X, Zhao J, Song W, Guo C, Li H. Nonlinear degradation modeling and maintenance policy for a two-stage degradation system based on cumulative damage model. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2016; 18 (2): 171–180, <http://dx.doi.org/10.17531/ein.2016.2.3>.
22. Ozkut M, Eryilmaz S. Reliability analysis under Marshall–Olkin run shock model. *Journal of Computational and Applied Mathematics* 2019; 349: 52–59, <https://doi.org/10.1016/j.cam.2018.09.022>.
23. Peng H, Feng Q M, Coit D W. Reliability and maintenance modeling for systems subject to multiple dependent competing failure processes. *IIE Transactions* 2010; 43: 12–22, <https://doi.org/10.1080/0740817X.2010.491502>.
24. Pourhassan MR, Raissi S, Hafezalkotob A. A simulation approach on reliability assessment of complex system subject to stochastic degradation and random shock. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2020; 22 (2): 370–379, <http://dx.doi.org/10.17531/ein.2020.2.20>.
25. Rafiee K, Feng Q M, Coit D W. Reliability modeling for dependent competing failure processes with changing degradation rate. *IIE Transactions* 2014; 46: 483–496, <https://doi.org/10.1080/0740817X.2013.812270>.
26. Rafiee K, Feng Q, Coit D W. Reliability assessment of competing risks with generalized mixed shock models. *Reliability Engineering & System Safety* 2017; 159: 1–11, <https://doi.org/10.1016/j.ress.2016.10.006>.
27. Segovia M C, Labeau P E. Reliability of a multi-state system subject to shocks using phase-type distributions. *Applied Mathematical Modelling* 2013; 37 (7): 4883–4904, <https://doi.org/10.1016/j.apm.2012.09.055>.
28. Tanner D M, Dugger M. Wear Mechanisms in a Reliability Methodology (Invited). *Proceedings of SPIE–The International Society for Optical Engineering* 2003; 4980: 22–40, <https://doi.org/10.1117/12.476345>.
29. Ye Z S, Tang L C, Xu H Y. A distribution-based systems reliability model under extreme shocks and natural degradation. *IEEE Transactions on Reliability* 2011; 60: 246–256, <https://doi.org/10.1109/TR.2010.2103710>.
30. Ye Z S, Xie M. Stochastic modelling and analysis of degradation for highly reliable products. *Applied Stochastic Models in Business and Industry* 2014; 31: 16–32, <https://doi.org/10.1002/asmb.2063>.
31. Zhang Y, Ma Y, Ouyang L, Liu L. A novel reliability model for multi-component systems subject to multiple dependent competing risks with degradation rate acceleration. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2018; 20 (4): 579–589, <http://dx.doi.org/10.17531/ein.2018.4.9>.
32. Zhao X, Wang S, Wang X, Cai K. A multi-state shock model with mutative failure patterns. *Reliability Engineering & System Safety* 2018; 178: 1–11, <https://doi.org/10.1016/j.ress.2018.05.014>.



Study of energy consumption of a hybrid vehicle in real-world conditions

Indexed by:



Jarosław Mamala^a, Mariusz Graba^{a,*}, Andrzej Bieniek^a, Krzysztof Prażnowski^a, Andrzej Augustynowicz^a, Michał Śmieja^b

^aOpole University of Technology, ul. Prószkowska 76, 45-758 Opole, Poland

^bUniversity of Warmia and Mazury, ul. Słoneczna 46A, 10-710 Olsztyn, Poland

Highlights

- The energy consumption of the vehicle is strongly dependent on the ambient temperature.
- The IC engine significantly increases the total energy expenditure of test cycles.
- CO₂ emissions from the PHEV's average fuel consumption are below the allowable limit.
- The use of a electric motor in vehicles significantly reduces the vehicle operation costs.

Abstract

The paper presents an analysis of energy consumption in a Plug-in Hybrid Electric Vehicle (PHEV) used in actual road conditions. Therefore, the paper features a comparison of the consumption of energy obtained from fuel and from energy taken from the vehicle's batteries for each travel with a total distance of 5000 km. The instantaneous energy consumption per travelling kilometre in actual operating conditions for a combustion engine mode are within the range of 233 to 1170 Wh/km and for an electric motor mode are within the range of 135 to 420 Wh/km. The average values amount to 894 Wh/km for the combustion engine and 208 Wh/km for the electric motor. The experimental data was used to develop curves for the total energy consumption per 100km of road section travelled divided into particular engine types (combustion/electric), demonstrating a close correlation to actual operating conditions. These values were referred to the tested passenger vehicle's approval data in a WLTP test, with the average values of 303 Wh/km and CO₂ emission of 23 g/km.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)

energy consumption, hybrid vehicle, road tests, energy share analysis.

Definitions/Abbreviations

\bar{a} - mean acceleration [m/s²],
 a_D - acceleration in the deceleration phase [m/s²],
 a - acceleration [m/s²],
AT - Automatic Transmission,
 C_F - Fuel consumption in test [kg/s],
 C_V - calorific value of fuel [J/kg],
 dV - speed change [m/s],
 E_M - energy consumption of motion [J],
EM - electric motor,
 E_T - total energy consumption [kWh],
 E_{Te} - total energy consumption of the electric drive [J],
 E_{Tf} - total energy consumption of the combustion engine [J],
EV - Electric Vehicle,
FD - free driving distance,
HEV - Hybrid Electric Vehicle,
ICV - Internal Combustion Vehicle,
 L - distance [m],
 L_A - distance of the acceleration phase [m],
 L_C - distance of the acceleration constant speed phase [m],

MT - Manual Transmission,
NUT - non -urban traffic distance,
 P - Power, [W],
PHEV - Plug-in Hybrid Electric Vehicle,
 Q_e - distance-based energy consumption of the electric drive [kWh/km],
 Q_f - distance-based fuel consumption of the combustion engine [dm³/100km],
 $Q_{T\text{ PHEV}}$ - distance-based energy consumption [Wh/km],
 Q_{Te} - total distance-based energy consumption of the electric drive [Wh/km],
 Q_{Tf} - total distance-based energy consumption contained in the fuel [Wh/km],
SOC - State Of Charge [%],
 T_L - time for stop phase or engine idle run [s],
 $t_{s,e}$ - start and end time of energy calculation [s],
 t_T - time traveled distance [s],
TTW - (Tank-to-wheels),
UT - urban traffic distance,
WLTP - The Worldwide Harmonized Light Vehicles Test Procedure,
 \bar{V} - average speed [m/s],

(*) Corresponding author.

E-mail addresses: J. Mamala - j.mamala@po.edu.pl, M. Graba - m.graba@po.edu.pl, A. Bieniek - a.bieniek@po.edu.pl, K. Prażnowski - k.praznowski@po.edu.pl, A. Augustynowicz - a.augustynowicz@po.edu.pl, M. Śmieja - smieja@uwm.edu.pl

V_c - speed in uniform motion [m/s],
 ΔE_D - energy losses of the internal combustion engine [J],

ΔE_E - energy losses of the electric drive [J],
 ΔE_L - energy losses by idle operating conditions of the vehicle [J].

1. Introduction

A passenger vehicle can be analysed in terms of the consequences of specific energy conversions occurring in its engine system. Combustion engines are the dominant engine type in most power train systems. As result of the energy conversions derived from the fuel delivered from the tank, the combustion engine generates heat energy which is then transformed into kinetic energy, transferred to the drive system and ultimately to the vehicle's wheels, thereby setting the vehicle into motion. In the energy balance of a moving vehicle, implementing a selected speed profile, the energy generated from the burnt fuel E_T is expended to drive the vehicle and also lost as result of various energy conversions occurring both in the engine and in the transmission system. Therefore, according to equation (1), it is a sum of the following: energy delivered by the drive system to the wheels and defined as the motion energy consumption (E_M) required for overcoming the vehicle's motion resistance, the drive system's energy losses (ΔE_E) and energy losses of the internal combustion engine (ΔE_D), as well as losses in energy by idle operating conditions of the vehicle (ΔE_L) including e.g. the vehicle's standstill phase:

$$E_T = E_M + \Delta E_E + \Delta E_D + \Delta E_L. \quad (1)$$

All components of the vehicle's energy balance vary over time and depend on the speed profile parameters and environmental conditions. A vehicle speed profile consist of 4 vehicle motion phases (accelerated motion, constant speed motion - constant speed, delayed motion, and standstill), the energy expenditure is estimated between start and stop of the vehicle and the their kinetic energy is equal to zero at the beginning and end. The description of the speed profile parameters, consist: average speed \bar{V} , travel distance L or average acceleration \bar{a} , is influenced by the share of particular profile phases i the given road section. The simple speed profile does not occur in practice. Complex speed profiles occur in reality, where the profile's kinematic parameters (speed, acceleration) are averages of many simple profile components (simple modules). The average values can be calculated from equations (3) and (4), where the average speed of a complex profile can be calculated from dependency [29]:

$$\bar{V} = \frac{\sum_i L_i}{\sum_i \int_{t_s}^{t_e} \frac{dv}{a} + \sum_i \frac{L_c}{V_c} + \sum_i \int_{t_s}^{t_e} \frac{dv}{a_D} + \sum_i T_L}, \quad (2)$$

wherein (i) is the number of simple profiles and the complex profile's average acceleration from dependency (6):

$$\bar{a} = \frac{\sum (\Delta V)_i}{2L}, \quad (3)$$

where

$$\Delta V = V_e^2 - V_s^2. \quad (4)$$

Standstill is an undesired motion phase, because the combustion engine's operation results in the generation of energy from burnt fuel, which is not collected by the transmission system. In such a case, the drive system's efficiency is equal to zero. In this context, "Stop&Go" systems started to be used in vehicles [5, 23, 44], which in principle stop the combustion engine during standstill. An additional advantage of this solution is the reduction of emissions of harmful substances and CO₂ contained in exhaust gases into the environment. The share of the standstill phase depends on the speed

profile and environmental conditions [37, 42]. In paper [10], the authors put emphasis on the analysis of the share of particular vehicle motion phases in a complex driving cycle in urban and non-urban traffic conditions. The authors demonstrated that over 20% of the acceleration phase is implemented with acceleration in the range of 0 – 1 m/s² and over 15% of the acceleration is in the range of 1 – 4 m/s² and usually amounts to over 5% of the total vehicle travel duration, i.e. the driving intensity is very important in terms of fuel consumption. In paper [13], Fontaras et al. focused on fuel consumption on the view of the dynamics, demonstrating a slight energy consumption increase of approx. 5% for non-urban driving and nearly 70% for urban driving. These differences mainly derive from two different vehicle speed profiles resulting from the average speed and driving dynamics. In paper [15], the authors dealt with the optimisation of the engine's load selection and the transmission ratio's selection strategies during acceleration of a an ICV (Internal Combustion Vehicle). A change in the driving dynamics by extending the acceleration time by 1s in the case of acceleration in the range of 0 – 30 km/h and by 2 s in the range of 0 – 40 km/h allows for reducing fuel consumption by more than 5%. The authors [6] analysed dynamic parameters of different vehicles. The analysis covered a broad spectrum of vehicles, starting with motorcycles, through passenger vehicles and ending with commercial vehicles, determining the acceleration values of 0.45 – 2.87 m/s² and the mean range of 0.2 – 0.82 m/s². The high variation in acceleration affects fuel consumption, which is subjective and depends on the road type, driving style and speed profile. In papers [1, 43], the authors noted the variation in driving styles with reference to the implemented driving cycle in actual traffic conditions. The increase in driving dynamics described in the paper causes an increase in fuel demand from 40% in non-urban to 45% in urban traffic. In paper [14], the authors pointed to the varying vehicle energy consumption in real-world conditions depending of its acceleration dynamics. Road tests demonstrated substantial discrepancies in the distance-based fuel consumption fluctuating between 12.44 and 31.8 dm³/100km on a ¼ mile section, depending on the acceleration dynamics and transmission ratio selection in the transmission system. The selected transmission ratios with lower values resulted in a reduced fuel consumption with an average drive system efficiency fluctuating between 19.38 and 24.6% which are tested on a vehicle with an ICE (Internal Combustion Engine) modern downsized powertrain

On the other hand, the authors of dissertation [12] emphasised the constant speed vehicle motion phase and designated the highest efficiency points for an ICE meeting the Euro 5 standard for specific driving speeds. It was indicated that for the various types of power train systems tested, the optimal speeds in terms of fuel consumption may range from 70 to 75 km/h. In this regard, the authors of a different dissertation [4] analysed the impact of various transmission systems and emphasised the AT and MT transmissions, for which the maximum efficiency point at constant speed of 70 km/h was designated at 24%.

However, in terms of fuel consumption, regardless of the motion phase testing and analysis, it is key to enable kinematic energy recovery in a vehicle accelerated in a delayed motion phase, where in most cases the energy is dispersed into the environment by the braking system. The introduction of the hybrid engine system HEV (Hybrid Electric Vehicle) was aimed at reducing the driving system's energy loss through energy recovery [27, 38, 48].

In a vehicle with a conventional engine system ICV, only 12–25% of the energy derived from fuel is consumed for motion in urban traffic conditions. Most energy is lost by the combustion engine in the form of emitted heat, own losses deriving from friction, and ineffective combustion in urban driving cycle, hybrid vehicles have 21-40%

of energy derived from fuel and electrochemical battery available for their disposal [11, 45].

In paper [46], the authors compared the combustion engine systems ICV with hybrid powertrain system (HEV) in terms of the driving style and demonstrated that the driving dynamics substantially affects the fuel consumption. In the ICV, the difference is as high as 74%, while in the HEV – 105%. In paper [24], the authors simulated various distance of a cycle consisting of the acceleration phase and the subsequent run-down phase in terms of reduction in fuel consumption in a hybrid electric vehicle. The results obtained demonstrate the potential to reduce fuel consumption depending on the speed range from 5 to 11% when applying an adequate acceleration intensity.

However, regardless of the engine type used, i.e. combustion or electric, or the interoperability of both as a hybrid powertrain system, the aforementioned environmental components affect the fuel consumption in actual operating conditions. In this paper, the authors emphasised the energy expenditure converted to the vehicle weight and distance for a modern PHE) used in various operating and traffic conditions. It is an modern powertrain with two energy storage units (fuel and batteries) and two drive units (ICE and EM) which drive the vehicle together. The drive system's energy consumption is analysed in terms of the TTW (Tank-to-Wheels), understood as the total expenditure of energy obtained from energy storage units referred to the distance travelled. The results were compared to the data obtained from the WLTP (The Worldwide Harmonised Light Vehicles Test Procedure) approval test.

2. Research on and development of hybrid electric vehicles

In the world around us, in which carbon dioxide emissions and environmental pollution are the main problem, electric vehicles are becoming increasingly popular. When compared to vehicles powered with petroleum derivatives, electric vehicles emit substantially less greenhouse gases and air pollutants. Thanks to technological progress, the operation of electric cars has become more user-friendly (e.g., increased mobility), mainly due to the improvement of energy storage parameters and optimization of energy consumption management by individual vehicle systems. Nearly all global car manufacturers are currently starting the development of entirely electrical models. On the other hand, customers are also attracted by the concept of using electric vehicles. The Allied Market Research (AMR) report [39], which provides a thorough analysis of the automotive market, reveals that technological advances and proactive government initiatives have led to an exponentially growing demand for fuel-efficient, low-performance, low-emission vehicles. The report also states that the increase in demand is fostered by strict exhaust gas emission regulations imposed in many countries. On the other hand, technological progress and proactive governmental initiatives ensure an exponential growth of the automotive market.

It is expected that in the next 30 years, the global production of new vehicles will increase by nearly 30% [14], resulting in the presence of over $2 \cdot 10^9$ vehicles on the Earth in several dozen years [3, 8, 17]. Due to the imperfections of currently produced vehicles, there is a need for continuous improvement of modern drives. Therefore, innovative solutions are implemented for the individual components of the vehicle, which will, on the one hand, increase mobility, and, on the other hand, contribute to the protection of the natural environment. New vehicles will be equipped with advanced drive systems with uniform or hybrid engines due to the introduction of increasingly strict standards on exhaust gases and carbon dioxide emissions [32]. It was announced that in 2025, the European Union will introduce a new exhaust fume emission standard named EURO 7, due to which meeting the new emission limits in uniform combustion engine systems will be very difficult or even impossible while maintaining high vehicle traction parameters related to the dynamics and average travel speeds [26]. However, regardless of the engine system used, battery

electric engines will be commonly used. The ion-lithium batteries used currently are quickly discharged and require frequent charging. The most novel changes in terms of battery weight reduction and performance improvement are lithium sulphur cells. They are fully compostable and biodegradable organic batteries that will not only be a good eco-friendly option, but also allow for rapid charging. They are also substantially lighter [16]. To allow batteries to easily meet the presented requirements, ultra-capacitors characterised with excellent parameters, especially at low temperatures, are added to vehicles. The ultra-capacitor's and lithium-ion battery's interoperability management requires using a hybrid energy storage system (HESS) with a suitably developed management strategy [47]. Currently, research is being carried out on the optimization of electric power supply systems, which include fuel cells [9, 25].

From the driver's point of view, the energy sources used are of no significant importance. In light of the requirements for a vehicle as an energy system, it is important to ensure adequate traction parameters capable of moving it in a satisfactorily short time on a given road section. In the current state of automotive development, the variety of hybrid or electric engine systems offered by manufacturers is broad, but their market share is insignificant. In the next 10 years, the dominant drive systems will most probably be the PHEV (Plug-in Hybrid Electro Vehicle). This is due to the fact that they combine the advantages of an electric motor with the energy autonomy derived from the limited range of EV (Electric Vehicles). Hybrid engine systems became dominated by such units as the combustion and electric engines, combined in parallel. This results from the greater universality of such an engine system solution in every-day use in urban and non-urban traffic [2, 6, 21, 34, 35]. The testing of hybrid engine systems powered with fuels are conducted with reference to the harmful component emission limits [18, 22, 28, 33, 35, 41, 42]. However, many authors are conducting tests of energy consumption in normal operating conditions [19, 20, 35, 36, 40] or solely with reference to the electric engine system [7, 35]. The real test constituting verification of such hybrid engine systems in terms of energy consumption are road measurements conducted in actual operating conditions. Therefore, this paper features an analysis of the impact of road conditions on the energy consumption in a hybrid engine system. For this purpose, a vehicle was tested on a distance of 5000 km, in three groups, with selected three travel distances:

- I – urban traffic (UT) with distance up to 20 km,
- II – non-urban traffic (NUT) with distance up to 70 km C,
- III – free driving (FD) with distance travels above 70 km D.

All road test was occurred for randomly selected drivers. All above mentioned speed profile parameters were recorded for each travel distance separately.

3. Research topic motivation

The difference in the energy value of energy carriers stored in passenger cars with hybrid drive systems means that a direct comparison of the mileage consumption for an internal combustion engine with the mileage consumption for an electric motor is not adequate in terms of unit. The use of the distance-based energy consumption in the standardized energy unit Wh/km for both drive units within the hybrid drive system allows to increase the possibility of their comparison. The comparative parameters may be the time of use of both drive units, energy expenditure and the possibility of relating the values obtained in operational tests to the values obtained in the approval test. The unit Wh/km adopted in this study is not compatible with the SI system, but it is used in the automotive industry and approval tests. Despite the similarity in the drive train for both drive units, the drive unit decides about the energy expenditure from energy storage. Thus, the motivation to undertake the research was the analysis of the energy parameters of the hybrid drive system for a given car trip, taking into account the drive unit used in real-world conditions. At the same

time, it was decided to examine the share of individual drive units in the vehicle's mileage consumption. Additionally, the analysis covered the influence of the ambient temperature on the electricity consumption and, as a result, the vehicle range.

4. Methodology

4.1. Distance-based energy consumption

The distance-based energy consumption is understood as the energy demand from the vehicle's energy storage units to its engine per travelled kilometre. In the case of the ICE, the total energy (E_{Tf}) can be formulated as a product of the fuel consumption (C_F) and the fuel calorific value (C_V):

$$E_{Tf} = C_V \cdot \int_{t_s}^{t_e} C_F dt, \quad (5)$$

where:

- C_F – fuel consumption [kg/s];
- C_V – fuel calorific value depending on the fuel's type [J/kg];
- $t_{s,e}$ – energy calculation start and end time [s].

For the electric motor unit, the total energy (E_{Te}) expended by the drive depends on the electric engine's structure, whether it is powered with direct or alternating current, and on the instantaneous output supplied from the batteries to the electric engine unit. In the case of the alternating current, the total energy can be calculated from equation (6):

$$E_{Te} = \int_{t_s}^{t_e} U(t)I(t)\cos\varphi(t) dt, \quad (6)$$

where:

- U - voltage over time,
- I - current amperage rating over time,
- $\cos\varphi$ - power factor (for direct current $\cos\varphi=1$),
- $t_{s,e}$ - start and end time of power take.

The total energy supplied to the vehicle's drive system in the case of a PHEV is the sum of the energy collected from various energy storage:

$$E_T = E_{Tf} + E_{Te}. \quad (7)$$

Energy recovery of the tested vehicle is not the subject of analysis in terms of operation, because it replenishes the energy storage unit by charging up batteries and thereby increasing the vehicle's travel range.

The total energy consumed by the vehicle per distance travelled represents the distance-based energy consumption, which can be compared to the values obtained in the WLTP test, expressed in Wh/km, following dependency:

$$Q_{T_PHEV} = \frac{E_T}{L}. \quad (8)$$

The obtained values vary and depend on the type of engine unit used and on the traction parameters: average travel speed, travel distance, and time.

4.2. Research program

The research concerned the analysis of the distance energy consumption in a selected passenger vehicle equipped with the Plug-in type hybrid engine system with consideration of the following:

1. analysis of the operating time of particular engine units in the hybrid engine system,
2. analysis of the total energy expenditure in instantaneous and incremental terms
3. analysis of the total distance-based energy consumption for the PHEV and consumption broken down into particular engine units,
4. analysis of the vehicle's range in different environmental conditions (temperature).

The traction and energy parameters were monitored using the Mercedes software for mobile devices and the TEXA diagnostic system, which allowed the recording of the following data: total vehicle range, divided into particular engine/motor, capacity of energy storage, total distance, distance for each drive units, travel time, mean speed and energy expenditure as the distance-based fuel consumption and distance-based energy consumption.



Fig. 1. Measurement system diagram

The aforementioned data was systematically recorded in the database and then analysed. The analysis of the distance-based energy consumption was conducted for the vehicle's actual operating conditions deriving from every-day travels divided into three groups. The travels were characterised by freedom in route selection and random selection of drivers with a standard hybrid engine system control mode. All tests were carried out with the battery fully charged (SOC = 100%).

4.3. Test and analysis of energy consumption

The distance-based energy consumption testing of an PHEV vehicle in actual operating conditions of the analysed vehicles was conducted using the Mercedes-Benz A 250e vehicle. It is a passenger vehicle manufactured in 2021 with a full hybrid engine system, wherein two engine units (electric and combustion) are installed on the front drive axis. The engine units interoperate with the 8 F-DCT transmission (Front –Double Clutch Transmission), wherein the drive is transmitted to the front wheels.

The tested vehicle's technical and structural parameters are presented in Table 1. Table 1 presents the average energy consumption for the electric engine system and the average CO₂ emission according to the WLTP test, which was taken from the approval certificate [31].

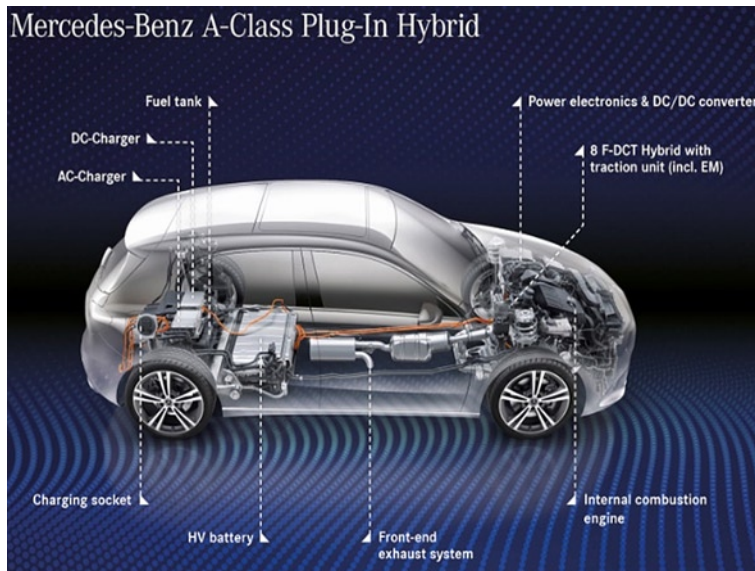


Fig. 2. Mercedes-Benz A-Class Plug-In Hybrid components [30]

Table 1. Tested vehicle parameters [31]

| | |
|---|--------------------------------------|
| Manufacturer | Mrecedes - Benz |
| Type | A250e / V177 |
| Combustion engine's displacement | 1332 cm ³ |
| Combustion engine's performance | 118 kW @ 5500 rpm |
| Combustion engine's max. torque | 210 Nm @ 1750 rpm |
| Electric engine's power | 75 kW |
| Long-term electric engine's power | 55 kW |
| Electric engine's max. torque | 300 Nm @ 0 - 5000 rpm |
| Engine assembly | Front, transverse |
| Combustion engine's supercharging | Supercharger |
| Engine system type | PHEV |
| Transmission system | Automatic - 8 gears |
| Battery capacity | 15.6 kWh |
| Vehicle weight | 1817 kg |
| Emission standard | Euro 6 (AP) |
| Travel range for petrol | 450 km |
| Travel range for batteries | 75 km |
| Average CO ₂ emission acc. to WLTP | 23 g/km (1.0 dm ³ /100km) |
| Energy consumption for the EV system | 209 Wh/km |

It is necessary to note the increase in the tested vehicle's weight in comparison to the internal combustion vehicle by nearly 300 kg due to using additional electric engine components (energy storage unit, electric engine, inverter and control system).

5. Test results

According to the adopted methodology, the study of the distance-based energy consumption in real-world cycles was conducted for the vehicle's actual operating conditions derived from the vehicle's every-day operation. The analysis of the hybrid engine system was conducted by using every-day vehicle travels in various atmospheric and road conditions, i.e. urban and non-urban traffic, in Opole and surrounding areas. The driver was free to use any driving technique. The travel distance was divided to three groups according to the meth-

odology. Table 2 presents the traction and energy parameters for the analysed travel groups.

Groups I and II were dominated by the vehicle's electric engine system (EV), in which the combustion engine unit was activated temporarily to increase the instantaneous speed or support the vehicle's intense acceleration on the road. In such situations, both units interoperated as a whole powertrain system. Figure 3 presents the percentage share of particular engine units in the tested vehicle.

In terms of particular percentage shares, the combustion engine unit's share was increasing from 6% in group I in 22% in group III, with an average value of 14% for all travels. The average values of distances in particular travel groups varied, as presented in Table 2. The highest differences can be observed in the energy expenditure expressed in Wh/km, which is presented in Figure 4.

The presented dependencies of the share of the distance-based energy expenditure per kilometre travelled for particular travel groups vary and depend on the time particular engine units were used. In all travel groups, despite the dominance of the electric engine unit powered from batteries, it is the combustion engine unit's use that substantially increases the total energy expendi-

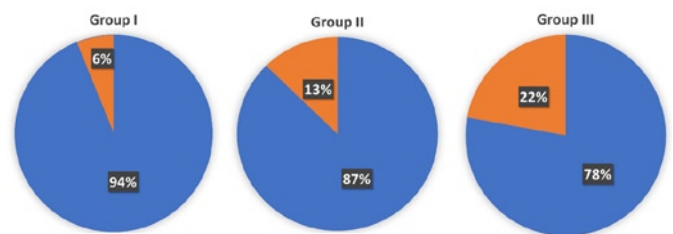


Fig. 3. Percentage shares of engine units for particular travels: a) group I, b) group II, c) group III (orange – combustion engine, blue – electric motor)

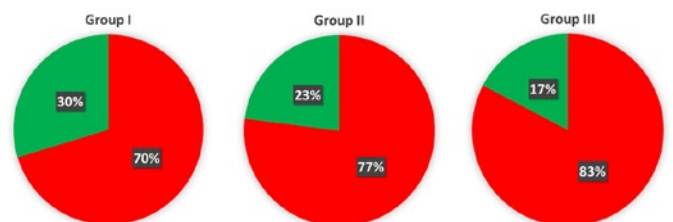


Fig. 4. Percentage share of particular engine units in the distance-based energy expenditure (green – combustion engine, red – electric motor)

ture in particular travels. When drawing the attention to the average values of the distance-based energy consumption Q_{T_PHEV} for all engine units in particular groups, it is possible to see that the values are higher than those deriving from the approval tests, which amount to 303.1 Wh/km for the tested vehicle. The travels in group II come closest to the above value, because the average distance-based energy consumption amounted to 356.2 Wh/km, which is 17.5% higher than the value achieved during the approval test. However, some travels carried out in groups I and II solely featured the use of the vehicle's electric engine system, the parameters of which are presented below:

In these terms, the average distance-based energy consumption achieved was lower than that achieved during the test. All mileages in the particular groups (Table 2) tested in the vehicle's actual operating conditions demonstrated substantial differences derived from driving the vehicle using particular engine units as well as substantial varia-

Table 2. Average engine system operating parameters during travels carried out in the hybrid engine system's standard operating mode

| Group | L_T [km] | L_e [km] | t_T [s] | V [km/h] | Q_f [dm ³ /100km] | Q_e [kWh/100km] | Q_{Tf} [Wh/km] | Q_{Te} [Wh/km] | Q_{T_PHEV} [Wh/km] |
|---------|---------------|---------------|--------------|---------------|-----------------------------------|----------------------|---------------------|---------------------|--------------------------|
| I | 9.1 | 7.7 | 1086 | 28.7 | 1.26 | 28.3 | 828.6 | 350.4 | 401.2 |
| II | 54.1 | 45.2 | 3935 | 51.8 | 1.10 | 18.2 | 740.6 | 216.4 | 285.2 |
| III | 121.9 | 83.8 | 8880 | 52 | 2.78 | 12.1 | 828.9 | 173.6 | 381.7 |
| Average | 51.4 | 41.7 | 3784.5 | 47.7 | 1.22 | 19.6 | 760.6 | 234.9 | 310.4 |

Table 3. Mean engine system operating parameters during travels carried out using solely electric engine unit

| Group | L_T [km] | L_e [km] | t_T [s] | V [km/h] | Q_f [l/100km] | Q_e [kWh/100km] | Q_{Tf} [Wh/km] | Q_{Te} [Wh/km] | Q_{T_PHEV} [Wh/km] | L_T [km] |
|-------|---------------|---------------|--------------|---------------|--------------------|----------------------|---------------------|---------------------|--------------------------|---------------|
| I | 7.1 | 7.1 | 960 | 25.2 | 0 | 33.35 | 0 | 333.5 | 333.5 | 7.1 |
| II | 46.1 | 46.1 | 2994 | 56.14 | 0 | 20.56 | 0 | 205.6 | 205.6 | 46.1 |

tion in the energy expenditure or distance-based energy consumption. It is difficult to directly compare the average distance-based fuel or energy consumption presented in Tables 2 and 3 in aspect to energy densities of the energy carriers in the storage units (Fig. 4a-c). Figure 6 presents the distance-based energy consumption for an approval test travel with reference to all travels for the PHEV hybrid system. When converted to energy expenditure derived from the used test vehicle, it is 25% higher than that achieved in the WLTP cycle at an mean speed of 13.25 m/s and mean distance of 51400 m (Table 2). This average parameter resulted from the actual road conditions correspond to the distance travelled by the test vehicle. This value was about 200% higher than that travelled in the WLTP test, wherein the traffic test amounts to 23266 m. It must be noted that the average vehicle speed was similar in the WLTP test and in actual operating conditions 12.92 m/s. The observed excessive distance-based energy demand for all travels (Fig. 5) exceeds the values recorded during particular travels above 140 km, which substantially exceeds the electric system's storage unit range. Therefore, in the case of travels in group II, which feature almost identical mean speeds and the distance was lower then energy storage unit's range. The distance-based energy

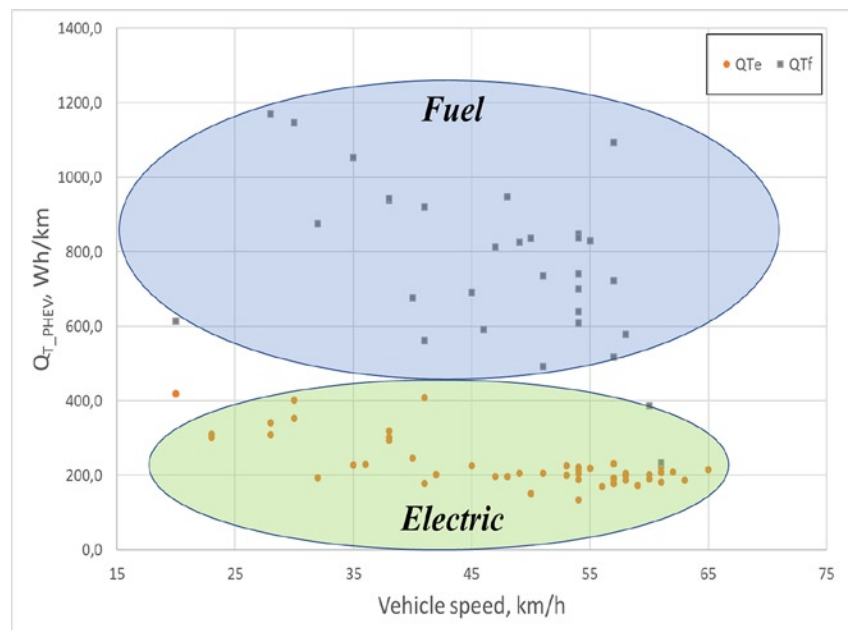


Fig. 6. Average speed and distance-based energy consumption broken down into particular engine units

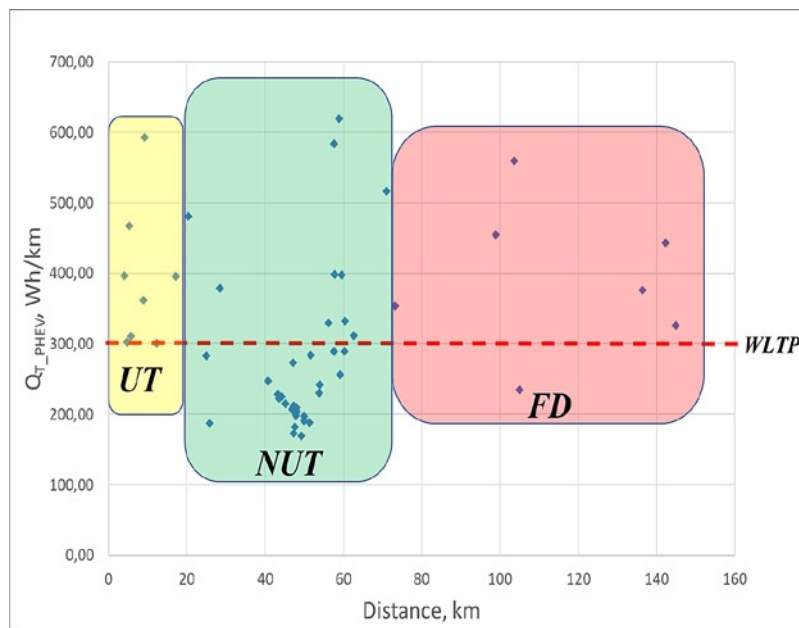


Fig. 5. Distance-based energy consumption refers to the distance travelled

consumption amounts to 285.2 Wh/km for the hybrid engine system and 205.6 Wh/km for the electric engine system and is 6% below the WLTP test value. The research result presented in Figure 5 were compared to WLTP homologation value, wherein the vehicle's average unit energy consumption were superimposed on particular travel groups. The travels carried out up to the energy storage unit's range do not exceed the distance-based energy consumption achieved in the WLTP test. The distance-based energy consumption in the case of the tests drives made from the electric energy storage does not exceed the values obtained during the WLTP approval tests (Fig. 5). In group III, there does not exceed the WLTP cycle value. However, this derives from a fast charging of the batteries during the test travel.

When drawing attention to the vehicle travel groups, the highest distance-based energy consumption was achieved in travels, during which the combustion engine system was used. The average instantaneous energy expenditure amounts to 760.6 Wh/km for the combustion engine system, i.e., approx. 2.86 MJ per kilometre travelled (Fig. 7). These values are 320% greater as the average energy expenditure for the electric motor. The

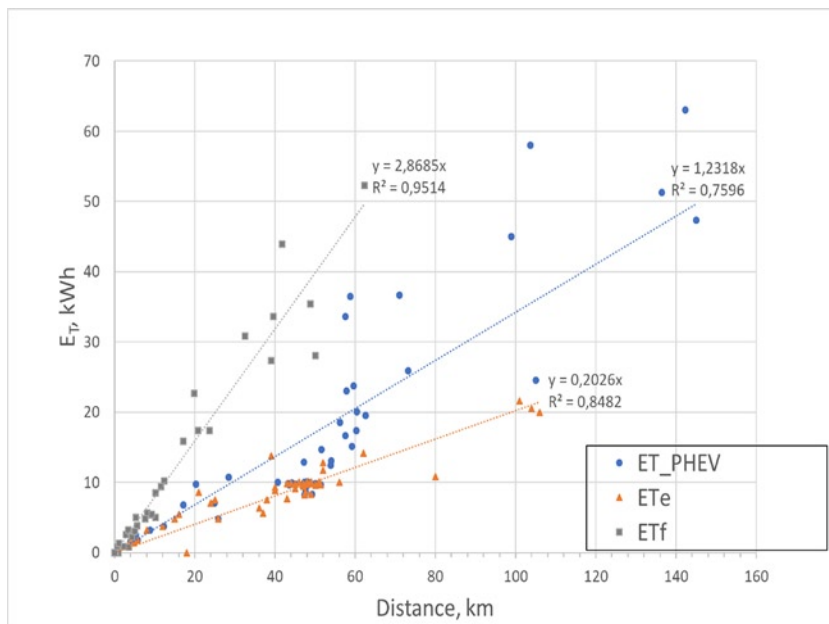


Fig. 7. Total energy expenditure to cover the given distance using various engine unit types with reference to the total distance travelled [29]

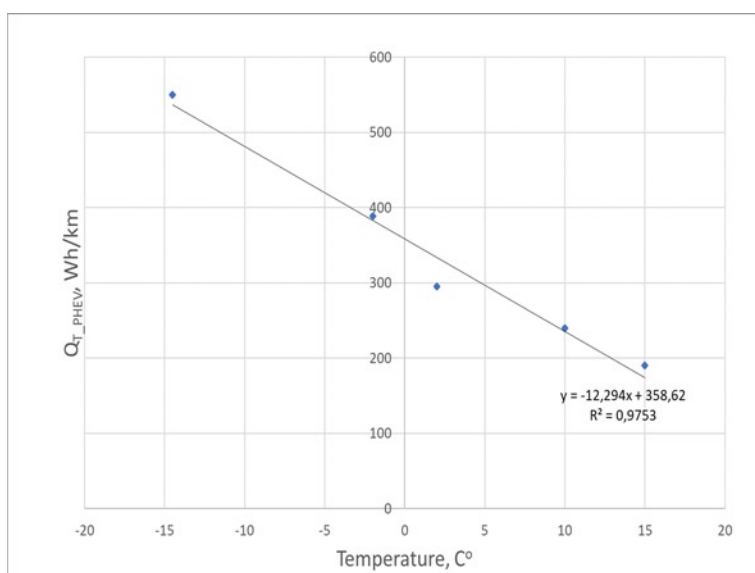


Fig. 8. Changes in the vehicle's total unitary energy consumption in different environmental conditions

average energy expenditure per kilometre travelled of which amounts to 234.9 Wh/km, which is equal to 0.72 MJ/km. This difference results mainly from the efficiency of the power units used [29].

In the case of city and highway driving, it can be expected as a significant increase in energy demand due to high dynamics or significant aerodynamic resistance. The value's decrease is more important when using the combustion engine unit, which results in more than a double reduction in the unit energy expenditure (from approx. 1500 Wh/km at an average speed of approx. 30 km/h to below 600 Wh/km at an average speed exceeding 60 km/h). In these terms, Figure 7 presents the total energy consumption for covering the given distance in terms of the total distance travelled and broken down into particular engine units used to drive the vehicle.

The research was based on the relation between the energy storage unit's capacity and ambient temperature. The issue of battery capacity reduction related to ambient temperature described in the literature was observed. Figure 8 presents an increase in average distance-based energy demand in the temperature range of -15 to 15°C . This constitutes another factor that results in the vehicle's reduced range. An increase in energy demand at low temperatures derives mainly from the additional energy expenditure to heat the interior and the battery assembly, but also from the increased motion energy consumption due to increased motion resistance.

It is necessary to note the vehicle's total unit energy consumption when using the electric engine, wherein the energy consumption at a negative temperature -15°C is over twice as high as at a positive temperature $+15^{\circ}\text{C}$. In this case, the vehicle's range was reduced by 21 km.

The designated straight line's regression coefficients for the vehicle's powertrain (Fig. 7) can be used for estimating the vehicle's operating indexes during the selected travel and road section. When calculating the energy expenditure, is it then possible to calculate the operating costs and the CO_2 emission different powertrain system. The mean energy consumption for a distance of 50000 m is presented in Table 5.

Attention must be drawn to the energy storage unit's capacity, which for the tested vehicle theoretically allows for covering a 75 km distance at the temperature of 18°C , after which the driver can only use the combustion engine. The tested hybrid vehicle allows for achieving the assumed data deriving from the conducted WLTP test for travel group II (Fig. 6). The PHEV powertrain is a very good solution not only in terms of energy expenditure, but also in terms of the CO_2 emission reduction. On longer routes, it is necessary to remember to replenish the energy storage unit, which lasts for the time depending on the available power grid. The tested vehicle's average charging time from 0 to 100% SOC are:

- for 220V charger (2.2 kW) – approx. 5.5 hours;
- for 380V charger (7.8kW) – approx. 1.5 hours;
- for CCS charger (22kW) – approx. 0.5 hours.

Table 5. Mean parameters during 50 km test distance carried out using the available hybrid system's operating modes [29]

| Type | L_T [km] | E_T [MJ] | Q_{T_PHEV} [MJ/km] | Q_f [dm ³ /100km] | Q_e [kWh/100km] | Price [Euro/100km] | CO_2 for TTW [g/km] |
|------|------------|------------|-----------------------|--------------------------------|-------------------|--------------------|------------------------------|
| EV | 50 | 36.3 | 0.72 | 0 | 20.2 | 2.87 | 0 |
| PHEV | 50 | 61.6 | 1.23 | 1.67 | 18.5 | 4.74 | 18.2 |
| ICV | 50 | 152.5 | 3.05 | 9.0 | 0 | 10.6 | 207 |

Table 6. Mean engine system operating parameters during travels carried out on a total distance of 5,200 km

| Distance [km] | dL_T [km] | L_e [km] | t_T [s] | V [km/h] | Q_f [dm ³ /100km] | Q_e [kWh/100km] | Q_{Tf} [Wh/km] | Q_{Te} [Wh/km] | Q_{T_PHEV} [Wh/km] |
|---------------|-------------|------------|-----------|------------|--------------------------------|-------------------|------------------|------------------|-----------------------|
| 5200 | 93.9 | 66.7 | 7692 | 41.7 | 2.78 | 14.75 | 893.7 | 208.6 | 410.1 |

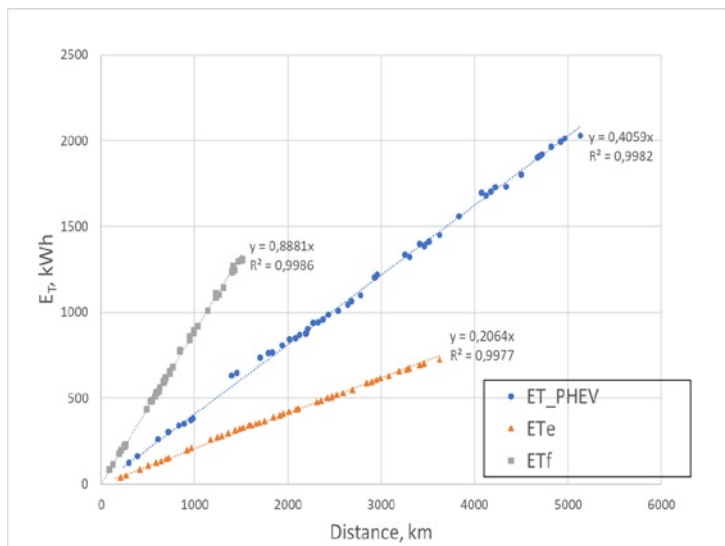


Fig. 9. Accumulated energy expenditure for particular hybrid system components

Using the manufacturer's data regarding the charging time, average energy price, and the obtained results of energy demand, it is possible to calculate the mileage costs. Assuming the average price of 1kWh of energy when using the power grid is 0.142 EUR/kWh and the unit price of energy in fuel (gasoline) is 1.18 EUR/dm³ (0.129 EUR/kWh), taking into account the energy/fuel consumption from individual storage tanks (battery, fuel tank) during the test driving's being the subject of the analysis (driving with the use of only electric drive, or driving only with the use of an ICE) causes a significant differentiation of operating costs (as energy costs) between the electric drive and the internal combustion engine. In this approach, the cost of energy consumed over a distance of 100 km for an internal combustion engine is approximately 3.7 times higher than for an electric motor (Table 5). The parameters of the drive system in terms of 5200 km the test cycle (Table 6).

The car total unitary energy demand over the distance of 5200 km (including trips using only the ICE) is greater compared to the results obtained only for the PHEV. It is related to the increase of ICE operating time up to 63.9%. These parameters were summed from the start of each distance of the test and counted from travel to travel as mean values. Despite to the greater PHEV vehicle's weight average fuel consumption amounts to 2.78 dm³/100km on a distance of over 5 thousand km. It means that the standards specified in the new regulations on CO₂ emission limit of 95 g/km from 2021 were met. When converted, the analysed vehicle's road emission amounts to 63.9 g/km and is below the acceptable limit.

An analysis of the distance-based energy consumption for urban and non-urban driving (travel groups I and II) demonstrate a double increase in energy consumption on short road sections (Fig. 5). When analysing the distance-based-energy consumption in increments presented in Figure 9, the parameter demonstrates a strong correlation of the energy expenditure to the distance travelled for particular engine systems. The obtained determination coefficient R^2 is equal to one and

the straight lines' direction coefficients changed slightly up to 4% in relation to the instantaneous values for particular travels.

6. Summary

The distance-based energy consumption of a passenger vehicle equipped with the Plug-in type hybrid powertrain in actual operating conditions presented in the paper presents a varied energy expenditure that depends on the engine unit used at the given time and driving cycle. The presented energy expenditure calculations based on standardised data for the tested vehicle allow for the formulation of conclusions in terms of the following:

1. Operating time of the hybrid drive system for individual drive units - in all groups of driving cycles, indicates the dominant electric drive unit (Fig. 3),
2. The energy expenditure per kilometre in instantaneous and increasing terms, shows a crucial increase in energy on the ICE (Fig. 4) and, divided by, generates more than a 4-times increase in the distance-based energy demand for the ICE compared to the electric motor in the TTW system,
3. The costs of energy consumption in real-world traffic conditions for the ICE are 3.6 times higher than for the electric drive (Table 5),
4. The range of a passenger car is consistent with the data given in Table 1, but under the condition of an appropriate ambient temperature of 18°C, in the conditions of an outside temperature of -15°C, the range has decreased almost four times.

The hybrid powertrain distance-based energy consumption in actual operating conditions for the analysed travel groups from I to III depends slightly on the average speed and driving style. The reference to the three groups of trips presented in the article, differing in terms of the traction parameters of the speed profile from the WLTP homologation test, enables their comparison after conversion to a standard unit of Wh/km. For driving in shorter distances than those resulting from the range of energy storage, the distance-based energy consumption is below the value obtained for the WLTP homologation test. This situation also applies to CO₂ emissions, which were recorded under operating conditions at the level of 38 g / km.

The indicators of the distance-based energy consumption of a passenger vehicle over a distance of 5200 km presented in the paper, in terms of average fuel consumption and estimated carbon dioxide emissions, are at a low level. The obtained value of road carbon dioxide emissions from average fuel consumption is 32.6% lower as the current standard in force from 2021.

In addition, the introduction of modern driver assistance systems in the test vehicle was also equipped, which makes a significant contribution to reducing fuel consumption and thus CO₂ emissions into the environment. An example is the navigation system, which affects the performance characteristics of the powertrain system, causing the drive system control algorithm to manage the energy consumption to the maximum extent to use the energy stored in the batteries on the route planned for navigation.

References

1. Barth M, Boriboonsomsin K. Real-world carbon dioxide impacts of traffic congestion. *Transportation Research Record* 2008; (2058): 163-171.
2. Barth M, Boriboonsomsin K. Energy and emissions impacts of a freeway-based dynamic eco-driving system. *Transportation Research Part D: Transport and Environment* 2009; 14(6): 400-410, <https://doi.org/10.1016/j.trd.2009.01.004>.
3. Becker T, Sidhu I, Tenderich B. Electric vehicles in the United States: a new model with forecasts to 2030. Center for Entrepreneurship and Technology, University of California, Berkeley, 2009: 36.
4. Bieniek A, Graba M, Hennek K, Mamala J. Analysis of fuel consumption of a spark ignition engine in the conditions of a variable load. *MATEC Web of Conferences*, 2017, <https://doi.org/10.1051/mateconf/201711800036>.

5. Bleek R. Design of a Hybrid Adaptive Cruise Control Stop- & -Go system. Engineering 2007.
6. Bokare PS, Maurya AK. Acceleration-Deceleration Behaviour of Various Vehicle Types. Transportation Research Procedia 2017; 25: 4733-4749, <https://doi.org/10.1016/j.trpro.2017.05.486>.
7. Chłopek Z. Research on energy consumption by an electrically driven automotive vehicle in simulated urban conditions. Eksploatacja i Niezawodność 2013; 15(1): 75-82.
8. Eder LV, Nemov VY. Forecast of energy consumption of vehicles. Studies on Russian Economic Development 2017; 28(4): 423-430, <https://doi.org/10.1134/S1075700717040049>.
9. Ehsani M, Gao Y, Emadi A. Modern Electric, Hybrid Electric, and Fuel Cell Vehicles. CRC Press: 2017, <https://doi.org/10.1201/9781420054002>.
10. Eisele WL, Turner SM, Benz RJ. Using Acceleration Characteristics in Air Quality and Energy Consumption Analyses Texas Transportation Institute The Texas A & M University System College Station, Texas 77843-3135 Southwest Region University Transportation Center Texas Transportation In. 1996.
11. Energy U S D of. Where the Energy Goes: Electric Cars. <https://www.fueleconomy.gov/FEG/atv.shtml> 2020.
12. Fontaras G, Franco V, Dilara P et al. Development and review of Euro 5 passenger car emission factors based on experimental results over various driving cycles. Science of the Total Environment 2014; 468-469: 1034-1042.
13. Fontaras G, Zacharof N G, Ciuffo B. Fuel consumption and CO2 emissions from passenger cars in Europe - Laboratory versus real-world emissions. Progress in Energy and Combustion Science 2017; 60: 97-131, <https://doi.org/10.1016/j.pecs.2016.12.004>.
14. Graba M, Mamala J, Bieniek A, Sroka Z. Impact of the acceleration intensity of a passenger car in a road test on energy consumption. Energy 2021; 226: 120429, <https://doi.org/10.1016/j.energy.2021.120429>.
15. He H, Cao J, Cui X. Energy optimization of electric vehicle's acceleration process based on reinforcement learning. Journal of Cleaner Production 2020; 248(ICEEE): 1-5.
16. Hong H, Che Mohamad NAR, Chae K et al. The lithium metal anode in Li-S batteries: challenges and recent progress. Journal of Materials Chemistry A 2021; 9(16): 10012-10038, <https://doi.org/10.1039/D1TA01091C>.
17. International Energy Agency. Energy Technology Perspectives 2017 - Executive Summary. 2017, https://doi.org/10.1787/energy_tech-2014-en.
18. Kitayama S, Saikyo M, Nishio Y, Tsutsumi K. Torque control strategy and optimization for fuel consumption and emission reduction in parallel hybrid electric vehicles. Structural and Multidisciplinary Optimization 2015; 52(3): 595-611, <https://doi.org/10.1007/s00158-015-1254-8>.
19. Kropiwnicki J. A unified approach to the analysis of electric energy and fuel consumption of cars in city traffic. Energy 2019; 182: 1045-1057, <https://doi.org/10.1016/j.energy.2019.06.114>.
20. Kropiwnicki J. Ocena efektywności energetycznej pojazdów samochodowych z silnikami spalinowymi. Wydawnictwo PG, Gdańsk 2011.
21. Kropiwnicki J, Furmanek M. Analysis of the regenerative braking process for the urban traffic conditions. Combustion Engines 2019; 178(3): 203-207, <https://doi.org/10.19206/CE-2019-335>.
22. Kum D, Peng H, Bucknor NK. Fuel and Emissions Reduction. Journal of Dynamic Systems Measurement and Control 2010; 2010(April): 1-18.
23. Kural E, Hacıbekir T, Güvenç B A. State of the art of adaptive cruise control and stop and go systems. arXiv 2020.
24. Lee J, Nelson D J, Lohse-Busch H. Vehicle inertia impact on fuel consumption of conventional and hybrid electric vehicles using acceleration and coast driving strategy. SAE Technical Papers 2009, <https://doi.org/10.4271/2009-01-1322>.
25. Li Q, Chen W, Li Y et al. Energy management strategy for fuel cell/battery/ultracapacitor hybrid vehicle based on fuzzy logic. International Journal of Electrical Power and Energy Systems 2012; 43(1): 514-525, <https://doi.org/10.1016/j.ijepes.2012.06.026>.
26. Limblici C. Investigation of engine concepts with regard to their potential to meet the Euro 7 emission standard using 1D-CFD software. 2020.
27. Liu T, Tang X, Wang H et al. Adaptive Hierarchical Energy Management Design for a Plug-In Hybrid Electric Vehicle. IEEE Transactions on Vehicular Technology 2019; 68(12): 11513-11522, <https://doi.org/10.1109/TVT.2019.2926733>.
28. Mamala J, Graba M, Praznowski K, Hennek K. Control of the effective pressure in the cylinder of a Spark-Ignition engine by electromagnetic valve actuator. SAE Technical Papers 2019, <https://doi.org/10.4271/2019-01-1201>.
29. Mamala J, Śmieja M, Praznowski K. Analysis of the total unit energy consumption of a car with a hybrid drive system in real operating conditions. Energies 2021, <https://doi.org/10.3390/en14133966>.
30. Mercedes-Benz. Mercedes me media. <https://media.mercedes-benz.com/>.
31. Mercedes-Benz. A250e homologation certificate. 2020: 1-30.
32. Merksiz J, Pielecha J, Radzimirski S. New Trends in Emission Control in the European Union. Cham, Springer International Publishing: 2014, <https://doi.org/10.1007/978-3-319-02705-0>.
33. Merksiz J, Rymaniak Ł. The assessment of vehicle exhaust emissions referred to CO2 based on the investigations of city buses under actual conditions of operation. Eksploatacja i Niezawodność - Maintenance and Reliability 2017; 19(4): 522-529, <https://doi.org/10.17531/ein.2017.4.5>.
34. Pielecha I, Cieślak W, Szalek A. Operation of electric hybrid drive systems in varied driving conditions. Eksploatacja i Niezawodność - Maintenance and Reliability 2018; 20(1): 16-23, <https://doi.org/10.17531/ein.2018.1.3>.
35. Pielecha I, Pielecha J. Simulation analysis of electric vehicles energy consumption in driving tests. Eksploatacja i Niezawodność - Maintenance and Reliability 2020; 22(1): 130-137, <https://doi.org/10.17531/ein.2020.1.15>.
36. Pitanuwat S, Sripakagorn A. An Investigation of Fuel Economy Potential of Hybrid Vehicles under Real-World Driving Conditions in Bangkok. Elsevier B.V.: 2015, <https://doi.org/10.1016/j.egypro.2015.11.607>.
37. Prochowski L. Movements Mechanics - Mechanika Ruchu. Warsaw, WKiŁ: 2016.
38. Qiu S, Qiu L, Qian L, Pisu P. Hierarchical energy management control strategies for connected hybrid electric vehicles considering efficiencies feedback. Simulation Modelling Practice and Theory 2019; 90: 1-15, <https://doi.org/10.1016/j.simpat.2018.10.008>.
39. Raport. Electric Vehicle Market - Global Opportunity Analysis and Industry Forecast, 2020-2027. Allied Market Research 2020: 256.
40. Rill G. Road Vehicle Dynamics: Fundamentals and Modeling - 1st Edition. CRC Press: 2011.
41. Schudeleit M, Küçükay F. Emission-robust operation of diesel HEV considering transient emissions. International Journal of Automotive Technology 2016; 17(3): 523-533, <https://doi.org/10.1007/s12239-016-0053-6>.

42. Siłka W. Energy consumption of car movement. *Energochłonność ruchu samochodu*. WNT: 1997.
43. Spalding S. RACQ Congested Roads Report : The Effects on Fuel Consumption and Vehicle Emissions Prepared by RACQ Vehicle Technologies Department. RACQ 2008; (07): 1-9.
44. Stanton NA, Dunoyer A, Leatherland A. Detection of new in-path targets by drivers using Stop & Go Adaptive Cruise Control. *Applied Ergonomics* 2011; 42(4): 592-601, <https://doi.org/10.1016/j.apergo.2010.08.016>.
45. Thomas J. Drive Cycle Powertrain Efficiencies and Trends Derived from EPA Vehicle Dynamometer Results. *SAE International Journal of Passenger Cars - Mechanical Systems* 2014; 7(4): 1374-1384, <https://doi.org/10.4271/2014-01-2562>.
46. Thomas J, Huff S, West B, Chambon P. Fuel Consumption Sensitivity of Conventional and Hybrid Electric Light-Duty Gasoline Vehicles to Driving Style. *SAE International Journal of Fuels and Lubricants* 2017, <https://doi.org/10.4271/2017-01-9379>.
47. Xiong R, Duan Y, Cao J, Yu Q. Battery and ultracapacitor in-the-loop approach to validate a real-time power management method for an all-climate electric vehicle. *Applied Energy* 2018, <https://doi.org/10.1016/j.apenergy.2018.02.128>.
48. Yeo H, Hwang S, Kim H. Regenerative braking algorithm for a hybrid electric vehicle with CVT ratio control. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 2006; 220(11): 1589-1600, <https://doi.org/10.1243/09544070JAUTO304>.

A fault tree-based approach for aviation risk analysis considering mental workload overload

Indexed by:



Haiyang Che^a, Shengkui Zeng^b, Qidong You^b, Yueheng Song^c, Jianbin Guo^{b,*}

^aSchool of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

^bSchool of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

^cScience and Technology on Complex Aviation System Simulation Laboratory, Beijing 100070, China


Highlights

- A fault tree-based risk analysis method considering MWLOL is developed.
- A MWLOL gate is proposed based on Multiple Resources Model.
- New logic relationships due to MWLOL are added to traditional FT through MWLOL gate.
- The new analysis method obtains more rational results validated by Accident Report.

Abstract

Many lives and aircrafts have been lost due to human errors associated with mental workload overload (MWLOL). Human errors are successfully considered in existing Fault Tree Analysis (FTA) methods. However, MWLOL is considered through Performance Shaping Factors indirectly and its information is hidden in FT construction, which is not conducive to analyze the root causes of human errors and risks. To overcome this difficulty, we develop a risk analysis method where Multiple Resources Model (MRM) is incorporated into FTA methods. MRM analyzes mental workload by estimating the resources used during performing concurrent tasks, probably including abnormal situation handling tasks introduced by basic events in FT. Such basic events may cause MWLOL and then trigger corresponding human error events. A MWLOL gate is proposed to describe MWLOL explicitly and add these new relationships to traditional FT. This new method extends previous FTA methods and provides a more in-depth risk analysis. An accident, a helicopter crash in Maryland, is analyzed by the proposed method.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

risk analysis, human error, mental workload overload, Multiple Resources Model, Fault Tree Analysis.

1. Introduction

Human errors (HEs), defined as that a human diverges from a normative plan or task [12], are regularly cited as the main causes of the majority of accidents in complex systems [3, 12, 25, 31]. Their pivotal role in aviation accident occurrence has been quantitatively pointed out in many studies: roughly 70% of all accidents in commercial aviation and 80% in general aviation [30]; more than 80% of helicopter accidents [4]. Pervasiveness of the HEs in accidents guarantees a requirement to investigate the causes of HEs to prevent future similar accidents [49].

In aviation, multitasking is prevalent in aviation [40], especially in abnormal situations [21]. HEs contribute to more than 70% of aviation accidents, and many of HEs can be attributed to workload [10]. During multitasks, a large number of cognitive resources such as attention, processing capacity, and multi-task performance [16] are required to complete assigned tasks, but the human has insufficient resources available to dedicate to the tasks [5]. Then, a high level of mental workload, or mental workload overload (MWLOL, i.e. the excessive levels of mental workload), occurs.

Due to the multi-dimensional characteristic of mental workload, Multiple Resources Model (MRM) [41] and Visual, Auditory, Cogni-

tive, and Psychomotor method (VACP) [19] are well known for workload prediction in aviation (e.g. [29, 42, 44, 52]). Wang et al. [38] propose a colored Petri net model based on MRM and VACP to predict mental workload. MRM and VACP claim that MWLOL occurs when the total demand for cognitive resources is beyond a threshold and pilot performance degrades [48]. Gore and Jarvis [9] suggest that when the cumulative demands of cognitive resources exceed an arbitrary threshold of 7, the operator will be at great risk of MWLOL.

With the development of technology in today's aircraft, pilots have to process a considerable amount of complex information [23]. Their attention often requires to be split between multiple information and the risk of MWLOL has increased [11]. The MWLOL can cause errors or delay information processing [5], and may reduce the vigilance and alertness of pilots with catastrophic effects [33]. Therefore, the MWLOL constitutes a key element in safety and reliability of complex man-machine systems. In aviation area, most of the accidents, especially those fatal ones, occurred due to high levels of mental workload of pilots [35, 51]. Many lives and aircraft of the United States Air Force have been lost due to errors made during periods of flight associated with MWLOL and task saturation [23]. This makes prediction and assessment of pilot mental workload a major issue in aviation

(*) Corresponding author.

E-mail addresses: H. Che - chehaiyang@buaa.edu.cn, S. Zeng - zengshengkui@buaa.edu.cn, Q. You - youqidong@buaa.edu.cn, Y. Song - samantha125@126.com, J. Guo - guojianbin@buaa.edu.cn

safety. Effective accident prevention should incorporate mental workload into risk analysis models.

Probabilistic safety assessment (PSA) is a comprehensive, structured methodology to identify and understand the risks associated with hazardous activities in complex systems [39]. It can identify potential accident scenarios, assess their likelihoods and consequences, and improve system safety and operation [20]. There are many PSA techniques, among which fault tree analysis (FTA) is one of the most prominent techniques [28] and is the most recognized and widely used [15]. The aim of FTA is to find the primary causes of accident causation utilizing a top-to-down method. The basic events of FT can be HEs, software or hardware failures, or environment events [6]. To analyze HEs and study human behavior in accident occurrence, many studies propose an analysis concept that combines FTA, Task analysis (TA) and human reliability analysis (HRA) methods [6, 53, 55]. FTA identifies the root causes of an accident, while TA analyses the way human perform tasks and how they interact with machines or other colleagues. These analysis methods are complemented by using one of HRA methods, such as ATHEANA (A Technique for Human Error Analysis), THERP (Technique for Human Error Rate Prediction), HEART (Human Error Analysis and Reduction Technique), CREAM (Cognitive Reliability and Error Analysis Method), and HEIST (Human Error Identification in System Tools). Doytchev et al. [6] combine FTA and TA to analyze an accident of Bulgarian Hydro power plant. In their analysis, HEs are analyzed by the combination method of TA and HEIST, through which details about HEs in a realistic situation are revealed. Zhou et al. [55] incorporate CREAM into FTA to analyze Liquefied Natural Gas carrier spill accidents, and estimate likelihoods of risks using Monte Carlo Simulation. Zhou et al. [53] propose a hybrid HEART method and incorporate it and TA to FT construction for risk analysis.

Although previous FTA methods successfully consider HEs based on the combination of TA and HRA, they ignore human mental workload or describe MWLOL through Performance Shaping Factors (PSF) indirectly, such as “number of simultaneous goals” and “available time”. In doing so, the MWLOL information is effectively hidden in the logical structure of the FT, and task scenarios causing high mental workload cannot be identified. Therefore, it is unable to play a role in qualitative analysis. In addition, HEs should be best viewed as a joint product of the interactions of humans with other aspects of the system (software, hardware, etc.) in a particular external context [22]. These FTA methods cannot describe the logic relationships among human error events and other basic events due to MWLOL in the process of man-machine interaction: basic events such as equipment failures may cause the system in an abnormal situation, then introduce a new abnormal situation handling task which is time-shared with current tasks, and finally MWLOL occurs and triggers the corresponding human error events. Therefore, to deeply analyze the root causes of human errors and accidents, the MWLOL should be considered and described explicitly in FT construction.

In this paper, we focus more on MWLOL and it is incorporated into FTA. A modified FTA method is developed based on aforementioned FTA methods combined with TA and HRA [6, 53, 55]. This new method also makes use of TA describing and analyzing how and when the human interacts with the system or colleagues in the system. TA can create a detailed picture of human involvement, including the concrete operations and plans. Plans determine which operations should be performed simultaneously. Based on TA, human error identification, analysis, and quantification can be implemented with HRA methods. Then a traditional FT can be constructed. To overcome the difficulty of considering and describing MWLOL explicitly in traditional FTA, we introduce MRM to build a MWLOL mechanism model and develop a new logic gate (i.e. MWLOL gate) to incorporate MWLOL into previous FTA methods. Such gate can represent how MWLOL occurs and what its effects are, and it may add the logic relationships among basic events due to MWLOL to traditional FT construction. The proposed method represents a major extension from previous

FTA methods and provides a more in-depth risk analysis. A case study of helicopter crash in Maryland On January 10, 2005 is used to illustrate the effectiveness of the proposed risk analysis method.

This paper is organized as follows. In Section 2, we introduce the MWLOL and its contributions to aviation accidents. Section 3 presents the background and basic concepts of risk analysis. In Section 4, the proposed methodology is presented, while in Sections 5 and 6 application of the methodology with results and discussions are provided. Finally, the conclusions of this paper is presented in Section 7.

2. Aviation accidents due to MWLOL

With the improvement of intelligence and automation during flight, the role of the pilot has changed fundamentally, from the operator and controller of the system to the supervisor and decision-maker [24]. The applications of advanced technologies has greatly reduced the pilot's physical workload in modern aviation. However, in some cases, advanced equipment actually increases the overall mental workload. Objectively, the cockpit has become a workplace with a high incidence of MWLOL because of the highly intensive information. Pilots need to collect more than 30 pieces of information within 10s before and after the takeoff of a Boeing 747. In another case, 675 special abbreviations and hundreds of warning signals are contained in three displays under the windshield of the F/A-18 Hornet Fighter cockpit alone [50]. Pilots need process the increasing information and the allowable time for decision decreases. Therefore, flying a plane is often a heavy mental workload task, especially in abnormal situations. The pilots must constantly acquire and process much information from their eyes, ears and other sensory organs to avoid accidents.

It has become a universal phenomenon that multiple tasks cause mental workload to exceed the mental ability of pilots, which is called MWLOL. The pilots' capacities of information processing are stretched with increased task demands. The occurrence of MWLOL has affected the performance of pilots seriously, which reduces the efficiency and safety of the system. For example, when a pilot performs dual tasks with MWLOL, s/he will become involved in her/his current situation of the primary task while forget to perform the secondary task [23]. Consequently, the information of the secondary task is not perceived, which usually lead to perception errors, information-processing errors and slow decision-making. These HEs due to MWLOL are frequently identified as a major cause of accidents [23].

A certain survey on the reasons for aviation accidents shows that 60%~80% of aviation accidents relate to human errors, most of which are caused by MWLOL [10]. As mentioned in introduction section, most of the accidents, especially those fatal ones, occurs due to errors associated with MWLOL [35, 51]. According to statistics, among the 81 flight-grade accidents in Civil Aviation Administration of China during the 15-year period of 1980-1994, 15 were caused by MWLOL [26].

Consequently, it is a major issue to analyze pilot mental workload in aviation risk analysis. Evaluating and improving the pilot's mental workload can be helpful in improving pilot performance and reducing the likelihoods of accidents.

3. Background of research methods

In the previous section, the importance and contributions of MWLOL to accident are demonstrated. This section covers the necessary background for understanding the proposed method of aviation risk analysis considering MWLOL. An overview of MRM, FTA, TA, and HRA is illustrated below.

3.1. Multiple resources model

MRM is developed by [40, 41], which are the main references used here. MRM can well interpret the occurrence of MWLOL and decrement of human performance caused by the interference between several concurrent tasks [40]. It has been widely used in workload

prediction and assessment in aviation (e.g., commercial aviation [2] and helicopter [8, 19]).

MRM holds the idea that humans have several separate limited and allocable mental resources. It provides a computational model to predict total interference between a time-shared pair of tasks, which is the sum of two components, a 4-dimensional demand component (i.e., resource demand) and a multiple resource conflict component (i.e., degree to which overlapping resources are required). The four dimensions, shown schematically in Fig.1, consist of (1) Information processing stages, referring to perception, cognition and response progress, (2) Processing codes, representing the spatial and verbal working memory codes, (3) Input modalities, containing the visual and auditory channels to allocate attention, and (4) Visual processing, dividing visual modality into focal and ambient vision [41].

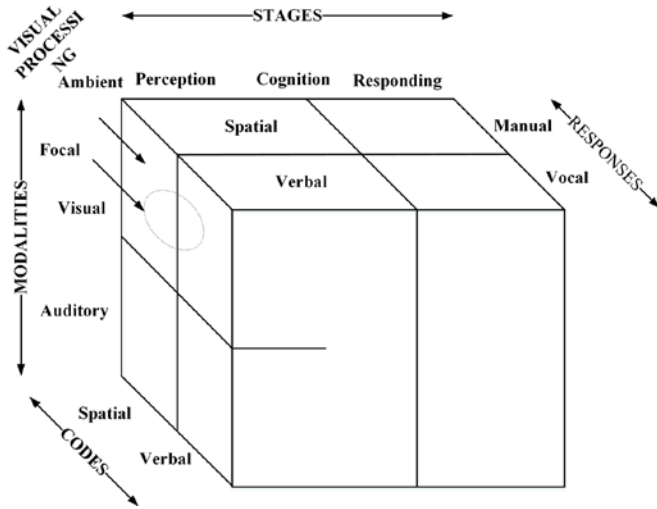


Fig. 1. The 4-dimensional MRM [41]

MRM evaluates task interference through the following three critical processes: (1) demand vector determination, (2) conflict matrix construction, and (3) total interference calculation [41].

(1) Basic mental resources demand reflects the mental workload to complete a single task. In MRM, the determination of resource demand value in certain dimension depends on the characteristic and difficulty of task. Each demand is specified as being automated ($d = 0$), easy ($d = 1$), or difficult ($d = 2$). According to the computational model, the demand vector of a certain task can be represented as: $\mathbf{d}_i = \{Vf, Va, As, Av, Cs, Cv, Rs, Rv\}$, where \mathbf{d}_i denotes the demand vector of task i ; V is visual; A is auditory; C is cognition; R is response; f represents focal vision; a represents ambient vision; s is spatial code; and v is verbal code. For the convenience of subsequent expression, the demand vector is simplified as: $\mathbf{d}_i = \{c_{i1}, c_{i2}, c_{i3}, c_{i4}, c_{i5}, c_{i6}, c_{i7}, c_{i8}\}$, where c_{ij} corresponds the j th ($j = 1, 2, \dots, 8$) elements in \mathbf{d}_i , and respectively represents the value of $Vf, Va, As, Av, Cs, Cv, Rs, Rv$.

(2) Based on plenty of studies, Wickens [41] proposed a conflict matrix to reflect the conflict value for different resource competitions intuitively, as shown in Table 1. If dual tasks use the same resources, the conflict extent will be the highest. Hence the dual tasks may be time-shared more easily when using different type of resources (e.g., perception vs. response, auditory vs. visual). The dual-task resource conflict score is determined by the summation of conflict values:

$$r(\mathbf{d}_1, \mathbf{d}_2) = \sum_{i=1}^8 \sum_{j=1}^8 \alpha_{ij} (c_{1,i} \otimes c_{2,j}), \text{ and } \alpha_{ij} = \begin{cases} 1, & c_{1,i} \neq 0, \text{ and } c_{2,j} \neq 0, \\ 0, & \text{else,} \end{cases} \quad (1)$$

where $r(\mathbf{d}_1, \mathbf{d}_2)$ denotes the resource conflict score between dual tasks T_1 and T_2 , and $c_{1,i} \otimes c_{2,j}$ is the conflict value of two resources, determined by Table 1.

Table 1. Conflict matrix proposed by Wickens [41]

| | | Task A | | | | | | | |
|--------|----|------------|-----|-----|-----|--------|-----|----------|-----|
| | | Perceptual | | | | Mental | | Response | |
| | | Vf | Va | As | Av | Cs | Cv | Rs | Rv |
| Task B | Vf | 0.8 | 0.6 | 0.6 | 0.4 | 0.7 | 0.5 | 0.4 | 0.2 |
| | Va | 0.6 | 0.8 | 0.4 | 0.6 | 0.5 | 0.7 | 0.2 | 0.4 |
| | As | 0.6 | 0.4 | 0.8 | 0.4 | 0.7 | 0.5 | 0.4 | 0.2 |
| | Av | 0.4 | 0.6 | 0.4 | 0.8 | 0.5 | 0.7 | 0.2 | 0.4 |
| | Cs | 0.7 | 0.5 | 0.7 | 0.5 | 0.8 | 0.6 | 0.6 | 0.4 |
| | Cv | 0.5 | 0.7 | 0.5 | 0.7 | 0.6 | 0.8 | 0.4 | 0.6 |
| | Rs | 0.4 | 0.2 | 0.4 | 0.2 | 0.6 | 0.4 | 0.8 | 0.6 |
| | Rv | 0.2 | 0.4 | 0.2 | 0.4 | 0.4 | 0.6 | 0.6 | 1 |

(3) The total interference value is represented by the sum of total resource demand value and $r(\mathbf{d}_1, \mathbf{d}_2)$:

$$TI = \sum_{i=1}^2 \sum_{j=1}^8 c_{i,j} + r(\mathbf{d}_1, \mathbf{d}_2), \quad (2)$$

where TI denotes the total interference of dual tasks.

3.2. Fault tree analysis

FTA is a well-established and well-understood technique, widely used to determine the causes of accidents and dig deep into the factors leading to these causes [14]. The analysis results allow practitioners to identify weaknesses in the system and take prevention methods. In this paper, the proposed risk analysis method considering MWLOL is implemented through FTA.

FTA is a top-down and graphical method that analyzes accidents deductively and structurally [55]. FTA starts with an undesired event as a top event usually representing the accident, and constructs downwards to dissect the system for further detail until the basic events leading to the top event are known [16]. The basic events are in the bottom of the tree, including human errors, mechanical failure, environmental factors and any other events that can caused accidents [6]. Their relationships are described by logic gates, such as AND-gate and OR-gate.

Once a FT is modeled, it can be analyzed in qualitative and quantitative ways [14, 46]. Qualitative analysis aims to find the minimal cut sets (MCS), which show how minimum basic events can combine together to cause the accident. In quantitative analysis, the probability of the accident occurrence and other quantitative indexes such as importance measures are mathematically calculated. The importance measures can determine which basic event in the cut sets are more critical to prevent the top event from occurring.

To capture the dynamic behavior of system failure mechanisms, the concept of dynamic FTA is proposed through adding the priority AND, standby or spare, and functional dependency gates to the traditional FTA [7, 47]. With the development of technology, many scholars have expanded the FTA to make them suitable for advanced and complex systems. Simultaneous-AND gate [37], AND-THEN gate [45], and SEQ-OR gate [18] are proposed to improve the modeling power of dynamic FTs.

3.3. Task analysis

Task analysis (TA) involves the study of the way operators perform the tasks in their work environment and how to refine these tasks into

a sequence of subtasks [6]. TA is the process of describing and analyzing how the operators interact with the system and other operators in order to achieve a system goal. TA can capture factors related to the cognitive activities of the human involved and psychological context of the tasks [1]. TA has experienced continuous improvement, and numerous TA methods have been developed, such as hierarchical task analysis (HTA), Goals Operations Methods (GOMS), Tabular Task Analysis, Timeline analysis, and cognitive task analysis [17].

Among the TA methods above, HTA is the “*best known task analysis technique*” [17], and has a very generic form that can almost be applied in any field. HTA focuses on the identification of the overall goal and the decomposition of the goal into subordinate goals and sub-tasks, which allows it to analyze complex tasks [1]. In HTA, the subordinate goals should be further decomposed into more detailed goals or tasks. Hence the decomposition needs to continue, until the sub-tasks in the bottom of HTA structure are all concrete operations. The goals and sub-goals are organized through plans, and the work processes are well structured based on its hierarchical approach [6]. The details and framework for conducting HTA can be seen in [32].

HTA has been extensively used in interface design and evaluation, allocation of function, job aid design, error prediction, and workload assessment [32]. In this paper, we focus on its application in the workload assessment and error prediction. These two parts deal with the question of how operators become MWLOL and human error occurs respectively. HTA is recognized as the pre-analysis before workload and human error analysis.

3.4. Human error analysis

As mentioned before, HEs are the main reasons for accidents in highly complex systems and the accidents caused by HEs has continuously increased [6]. Therefore, drilling down the causes of HEs is significant for accidents analysis. Human reliability analysis (HRA) is a series of techniques for human error analysis. The present HRA methods are almost based on the human factors engineering, mental science and probability statistics [55]. They aim at eliminating accidents attributed to HEs. To consider the impact of human errors, HRA methods usually include several stages i.e., decomposing human act, identifying error modes, calculating human error probability, determining effects and analyzing the reasons for HEs [53].

After decades of development, some classic HRA methods are gradually promoted, e.g., THERP (Technique for Human Error Rate Prediction), HEART (Human Error Assessment and Reduction Technique), CREAM (Cognitive Reliability and Error Analysis Method) [13], and etc. Among them, CREAM focuses more on cognitive error and holds the concept that the performance is mainly influenced by the context. Based on this concept, nine Common Performance Conditions (CPCs) are defined to represent how context, including environment, equipment, organization, and etc., influences the performance of operators in system. The influence level is divided into three categories, i.e. improved, reduced and insignificant levels.

CREAM classifies cognitive functions into four categories: observation, interpretation, planning, and implementation. Each category contains several failure modes, and each failure mode has its corresponding failure probability named Cognitive Failure Probability (CFP). CPCs can be utilized to calculate the CFPs and determine the causes of them. On the one hand, CPCs combine with basic probability to determine the fixed CFPs [55]. On the other hand, CREAM defines the causal relationship between CPCs. According to the causal chain, the causes of human errors can be traced. In this way, the contribution of MWLOL can be indirectly analyzed with CPCs like “number of simultaneous goals” and “available time” [13].

4. Methods

A brief overview of methods and techniques for risk analysis were introduced in section 3. The FTA, TA, and HRA methods focus either on the failure of machine or human, and their combinations are uti-

lized to analyze the causes of human errors and accidents. However, the main cause of pilots’ errors, MWLOL, was ignored or considered indirectly through CPCs. To better analyze the MWLOL and its effects, MRM is introduced and combined with TA as a means to identify time-shared tasks and their resource demands that prompt MWLOL. In addition, to analyze the way MWLOL leads to accidents, the proposed methods are complemented with the utilization of FTA and a new logic gate i.e. MWLOL gate. Through this gate, a new dependence among basic events due to MWLOL can be analyzed.

4.1. Procedure

The analysis flow is shown in Fig. 2 and consists of 8 steps. Traditional FT is first constructed with HTA and CREAM in steps 1-4. Then it is modified by a MWLOL gate to analyze MWLOL and corresponding effects. Accordingly, main steps are explained as follows.

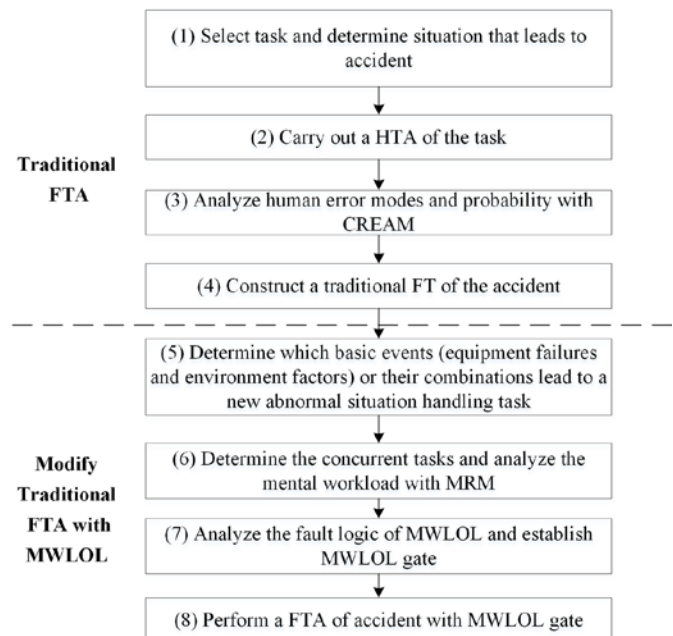


Fig. 2 Flowchart of the analysis approach

Step 1- Task selection and Situation determination: The tasks referring to flight handbook are the definition of steps that pilot must complete during a flying process. The purpose of the Situation determination is defining a variety of instant conditions according to the tasks assessed, such as working environment, task status, time availability and so on.

Step 2- Task analysis: In accordance with the situation, task analysis is carried out through HTA, and a list of subtasks is obtained. Multitasking is prevalent in aviation [40], and the majority of MWLOL occurs by performing Multiple tasks concurrently. Therefore, in this step, it is essential to determine the plans of these subtasks to identify which are time-shared.

Step 3- Human error analysis: CREAM is introduced to identify and analyze human errors in flying operations based on the results of TA. According to the historical data collected by National Transportation Safety Board (NTSB) or expert judgments, the cognitive function, CPCs and their weights can be obtained. Finally, the probabilities of human errors in pilots’ flying tasks can be calculated through CREAM [13].

Step 4- Perform a traditional FTA of the accident: There are many causes that can lead to the accident, such as equipment/mechanical failure, human errors, and environmental factors. Each of such causes is connected by logic gates and lower events until all its branches are terminated with basic events. Various logical combinations that lead to the accident can be displayed. Then FT is constructed and FTA is implemented to identify the root causes of the accident.

Steps 1-4 are the procedures to perform a traditional FTA, which is widely studied in many literatures [6, 53, 55]. To consider MWLOL, Steps 5-8 are proposed to modify traditional FTA in this paper, and we present them in sections 4.2-4.5 in detail.

4.2. Abnormal situation handling tasks caused by basic events (Step 5)

As the basic events including equipment/mechanical failure, human errors, and environmental factors occur, the aircraft may be in an abnormal situation. Pilots need deal with the abnormal situation based on the emergency procedure in flight handbook to prevent the accident [36]. Therefore, a new abnormal situation handling task is introduced, which will increase mental workload significantly. Then the performance of pilots will be affected seriously and the efficiency and safety of the system may be reduced. For example, single engine fire will introduce the engine fire extinguishing task. Pilots need perform at least dual tasks (i.e. flying task and extinguishing engine fire task) simultaneously. On such condition, the MWLOL may occur during man-machine interaction and lead to human errors and aircraft crash with high probability. Many aviation accidents have occurred when pilots perform multiple tasks besides an abnormal situation handling task.

In step 5, the abnormal situation handling tasks introduced by the occurrence of basic events are determined. Whether current tasks lead to MWLOL and what their effects are will be analyzed in steps 6 and 7 respectively.

4.3. Mental workload analysis with extension of MRM (Step 6)

Based on the results of HTA of normal tasks in step 2 and abnormal situation handling tasks in step 5, the tasks that should be performed concurrently are determined first in this step. Then mental workload analysis of these concurrent tasks is conducted with the extension of MRM.

In literature, MRM is proposed to predict the time-shared task interference, which is the sum of resource demands and conflicts. It is a convenient way to calculate mental workload caused by dual tasks. However, for the calculation of resource conflicts, it cannot be applied directly to the task scenario which contains three or more concurrent tasks. In aviation, especially under abnormal conditions, it is a common phenomenon that pilots perform multiple concurrent tasks [40]. To calculate multi-task interference, the above basic MRM is extended based on the following principles that the resource conflict is calculated according to task priority. For example, to calculate the resource conflict of three time-shared tasks, the resource conflict between the first and second highest priority tasks is first calculated, and then we calculate the resource conflict between the first two tasks and the third highest priority tasks. The detail steps are as follows:

First, tasks are ranked in descending order of priority based on TA. Let T_i denote the task with prioritization i . The prioritization of T_i is higher than that of T_{i+1} .

Second, let \mathbf{D}_p denote the sum of demand vector of T_1, T_2, \dots, T_p . It can be obtained through:

$$\mathbf{D}_p = \sum_{i=1}^p \mathbf{d}_i, \text{ and } 1 < p \leq n, \quad (3)$$

where n is the number of concurrent tasks, and $n > 1$.

Third, the resource conflict value between T_p and other tasks whose prioritization higher than T_p can be calculated with:

$$r(\mathbf{D}_{p-1}, \mathbf{d}_p) = \sum_{i=1}^8 \sum_{j=1}^8 \alpha_{ij} (c_{p-1,i} \otimes c_{p,j}), \quad (4)$$

where $c_{p-1,i} \in \mathbf{D}_{p-1}$, $c_{p,j} \in \mathbf{d}_p$. Then, the resource conflict score of multiple time-shared tasks can be represented as:

$$R(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n) = \sum_{p=2}^n r(\mathbf{D}_{p-1}, \mathbf{d}_p). \quad (5)$$

Finally, the total interference value of these tasks can be calculated as:

$$TI = \sum_{i=1}^n \sum_{j=1}^8 c_{i,j} + R(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n). \quad (6)$$

In summary, if the pilot need perform dual tasks concurrently, the mental workload analysis can be conducted by the basic MRM introduced in section 3.4, while if the pilot need perform three or more tasks concurrently, the mental workload analysis can be conducted by the extension of MRM proposed in this section.

4.4. MWLOL gate (Step 7)

For the contribution of MWLOL to aviation accidents, considering it into FTA model is beneficial to analysis the causes of accidents. The functions of FTA are reflected in various logic gates which represent how failures in subsystems can combine to cause a system failure. Therefore, it is a feasible method to construct a new logic gate (i.e. MWLOL gate) to model the fault logic of MWLOL. Based on the task management theory [43], tasks are abandoned in the order of priority when concurrent tasks lead to MWLOL.

4.4.1. Fault logic of MWLOL

“Mental workload describes the relation between the (quantitative) demand for resources imposed by a task and the ability to supply those resources by the operator” [41]. To investigate the fault logic of MWLOL, it is important to understand the strategy of task management that operators adapt when the supply is less than the demand. At a most general level, there are four possible types of adaptation when the MWLOL occurs [43].

- Operators may allow tasks' performance to degrade, for example, a vehicle driver may allow lane position to wander when the workload of dealing with an in-vehicle automation system increases.
- Operators may perform the tasks through a less resource consuming and more efficient way, as they may shift from optimal algorithms to satisfactory heuristics in decision making.
- Operators may shed tasks altogether, in an “optimal” fashion, eliminating performance of those of lower priority. For example, the air traffic controllers with mental workload overload may cease to offer pilots weather information unless requested, while turning their full attention to traffic separation.
- Operators may shed tasks altogether, in an “non-optimal” fashion, abandoning those that should be performed. For example, a vehicle driver abandons safe driving in favor of a cell phone conversation.

Unfortunately, beyond the studies and literatures on task management and resource allocation, very little is known about general principles that can account for when people adopt one strategy or the other [43]. However, training can certainly help operators to adopt an “optimal” strategy [43].

In this paper, the pilots are assumed to be well-trained, and they may shed tasks altogether in an “optimal” fashion, i.e., pilots under high workload will focus on the critical tasks with higher priority and eliminate performance of tasks of lower priority. Therefore, some of the operations for tasks of low priority will be abandoned.

MRM assumes that humans have several separate allocable mental resources but limited. Gore and Jarvis [31] suggest an arbitrary threshold of 7, i.e., the maximum cumulative demands of cognitive

resources people can provide is 7. Then whether the MWLOL occurs can be determined based on the value of total interference of time-shared tasks. Therefore, when the total interference of concurrent tasks exceeds 7, MWLOL is assumed to occur and the operator tends to eliminate performance of lower priority tasks.

4.4.2. Establish MWLOL Gate

As discussed above, the fault logic of MWLOL is obtained. We describe this fault logic by introducing a MWLOL gate, as shown in fig. 3. MWLOL Gate has multiple inputs and outputs. The inputs are concurrent multiple tasks (i.e. T_1, T_2, \dots, T_n), while the outputs (t_1, t_2, \dots, t_n) are the abandonment of tasks of low priority which triggers the corresponding human errors whose modes are omissions, such as perception omissions due to the abandonment. All inputs are basic events. Each output event t_i represents the abandonment of input event T_i . When all inputs occur simultaneously and the task interference exceeds 7 (i.e., the MWLOL occurs), the output events (t_n, t_{n-1}, \dots, t_2) will occur in turn until the interference of performing tasks is less than 7.

The MWLOL occurs only when multiple tasks need to be handled at the same time. Therefore, the output events occur only when all input events occur simultaneously. However, whether the input events of the AND gate occur at the same time or not, is not clear from its definition. The AND gate has no time parameters. To consider the temporal relations among input events, many logic gates have been developed to extend the description and analysis of fault trees, such as Priority-AND gate [7], AND-THEN gate [45], and Simultaneous-AND gate [37]. A Simultaneous-AND gate represents the input event X and Y occur at the same time. MWLOL Gate is proposed based on Simultaneous-AND. In this paper, the temporal relation that all input events occurs simultaneously is ensured by TA. The occurrence of output events depends on the MWLOL judged by MRM.

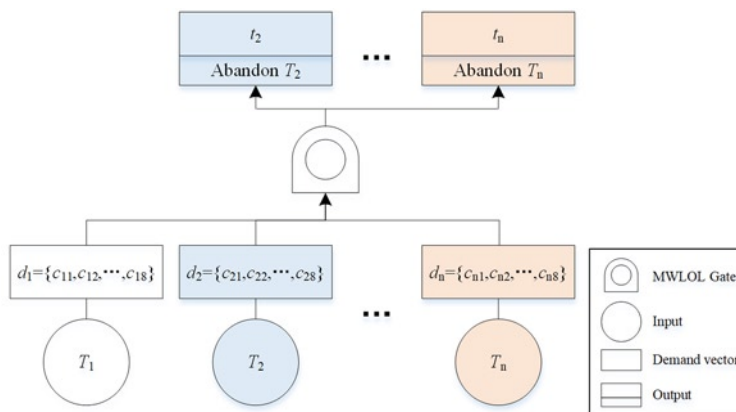


Fig. 3. Images of MWLOL Gate

According to the MRM and task prioritization strategies, the following rules of MWLOL GATE are made:

- 1) MWLOL GATE will not be triggered if only one input event occurs;
- 2) All output events will not occur if no MWLOL occurs;
- 3) All input events must be time-shared;
- 4) Output event t_i occurs later to t_{i+1} .

4.5. Modelling FTA with MWLOL gate (Step 8)

Based on Steps 1-3, task selection and situation determination, task analysis, and human error identification and analysis have been conducted. A combination of TA and HRA is utilized to determine the human error modes and their probabilities. Then, in step 4, a traditional FT can be established as shown in Fig. 4, and the detailed procedure can be seen in [6, 53]. The traditional FT considers human errors, ma-

chine failures, and environment factors. The top event (i.e. accident or incident) will occur when the MCS of basic events occur.

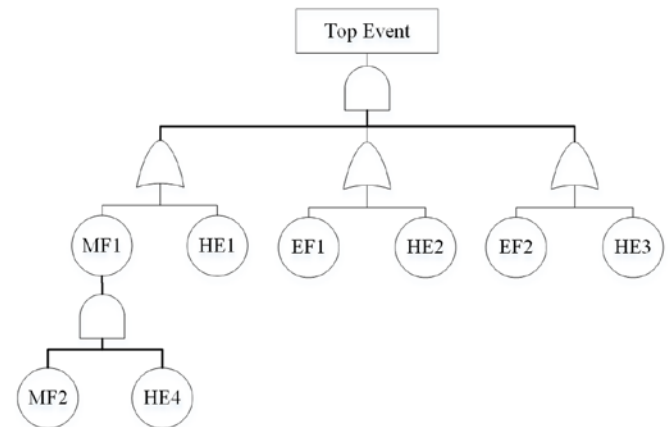


Fig. 4. Example of traditional FT with human errors (HE), machine failures (MF), and environment factors (EF) events

Based on steps 5 and 6, concurrent tasks (i.e. normal tasks and abnormal situation handling tasks) in the risky task scenario are identified and task interference can be calculated by MRM to identify whether MWLOL occurs. Step 7 establishes the MWLOL Gate that can determine which task will be abandoned when MWLOL occurs. Then its actual contents of output events trigger omission error events due to the abandonment of tasks. Therefore, for these omission error events, their occurrence is due to the MWLOL or omissions. Then, such omission error events in Fig. 4 will change from basic events to intermediate events, which are connected by OR gate and basic events t_i and omission, such as HE2 and HE3 in Fig.5. The probability of basic event “omission” can be calculated using CREAM.

MWLOL Gate represents how MWLOL occurs and what its fault logic is. By using MWLOL Gate, the MWLOL is present in the logical structure of FT, which plays a significant role in qualitative analysis of the root causes of aviation accidents.

Figure. 5 shows an example of a FT with MWLOL Gate. On such situation, operator need handle three tasks (i.e. T_1, T_2, T_3) simultaneously. Among them, T_3 is assumed to be the abnormal situation handling task caused by the basic events MF2 and EF1. The task interference of these three tasks exceeds 7, and the event t_3 (i.e. abandon T_3) occurs. Then t_3 triggers event HE3. In addition, if the task interference of T_1 and T_2 also exceeds 7. The event t_2 (i.e. abandon T_2) also occurs and triggers HE2. On the contrast, if the task interference of T_1 and T_2 is less than 7, the event t_2 will not exist, and HE2 is only affected by operation omission. Through MWLOL Gate, the dependence among basic events MF2, EF1, HE2, and HE3 can be described explicitly, and the causes of HE2 and HE3 can be well explained.

As shown in Fig. 5, MWLOL Gate combined with other logic gates can describe how the basic events cause top event. The causes of HEs in the process of man-machine interaction can be well investigated through the modified FTA. Then, HEs caused by MWLOL or not, mechanical failures and environmental factors can be identified as the root causes of accidents. Moreover, quantitative analysis like the calculation of top event probability and probability importance of basic events can be used to prioritize those causes. Therefore, FTA with MWLOL Gate can be analyzed in qualitative or quantitative methods, which are the same as traditional FTA.

5. Case study

The proposed analysis method is implemented to an accident of helicopter crash in Maryland [27]. On January 10, 2005, about 23:11, a helicopter crashed into the Potomac River during low-altitude cruise flight near Oxon Hill, Maryland. The pilot and several crews were

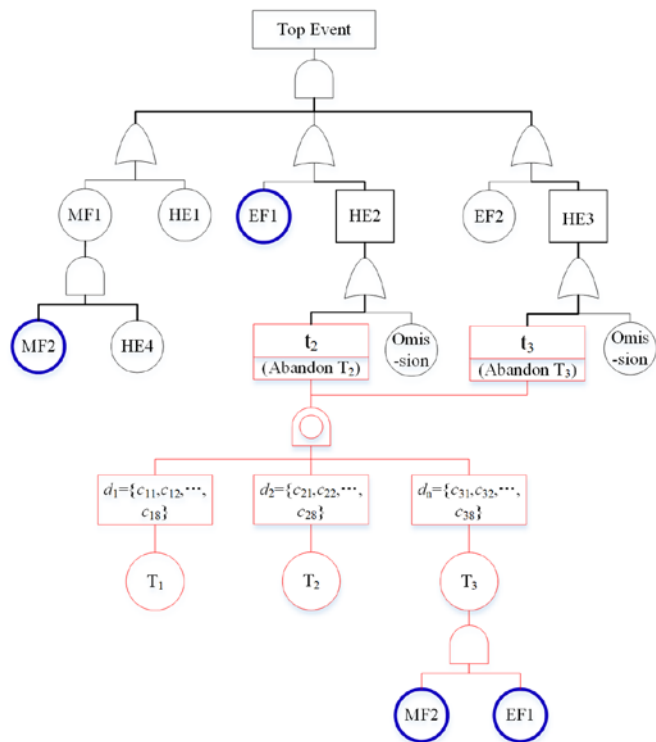


Fig. 5. Example of FTA with MWLOL Gate

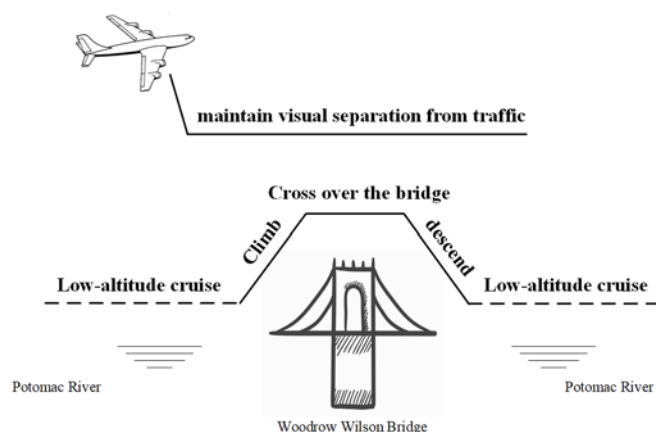


Fig. 6. Flight mission profile of helicopter

killed and the flight nurse was seriously injured. The helicopter (i.e. Eurocopter EC-135 P2, N136LN) was destroyed. The positioning flight was conducted under the provisions of visual flight rules (VFR) with a company flight plan filed.

The helicopter originated at the Washington Hospital Center Helipad and was en route to Stafford Regional Airport. During the flight route, multiple tasks besides an abnormal situation handling task should be performed simultaneously. Researching the MWLOL is worthwhile to analyze the root causes of the accident. The risky task scenario is analyzed using the proposed risk analysis method to illustrate its effectiveness.

5.1. Task selection and Situation determination

In view of the aviation accident report, the flight mission profile is as shown in Fig. 6. Since the flight route is near Stafford Regional Airport, the helicopter need cruise with low altitude and is usually asked to avoided airbus. The helicopter flies southbound along the Potomac River toward Woodrow Wilson Bridge. When the helicopter is near the bridge, climbs and crosses over the bridge. Then the helicopter descends and cruises with low altitude. During performing the tasks

above, the pilot is informed that an airbus was ten miles above the helicopter. The pilot should search for airbus visually and maintain visual separation from the airbus. Therefore, on such condition, the pilot need perform dual normal tasks concurrently, which may lead to MWLOL of the pilot. Considering the contribution of MWLOL to aviation accidents, these tasks are selected and analyzed in this section.

The situation can be determined based on a closer look at the aviation accident report. The pilot holds a commercial pilot certificate with ratings for airplane single- and multi-engine land, rotorcraft helicopter, and instrument helicopter. He is well trained and experienced. The helicopter was manufactured in 2004 and had accumulated 166.6 total flight hours at the time of the accident. The helicopter was configured one pilot, one flight paramedic, and one flight nurse.

The tasks are performed at night, about 23:11. According to the aviation accident report, a new moon was below the horizon and no illumination was provided at the time and location of the accident. Flying low-attitude North of the bridge is typically flying VFR due to the intense amount of ground lights available along the river. Once the pilot crosses the bridge he is now flying into a black void, and there is no outside visual reference. Therefore, the helicopter likes flying into actual instrument meteorological conditions, and flight instruments should be used to a greater degree to ensure altitude awareness.

5.2. Task analysis using HTA

Based on the helicopter flight handbook [36], with four raters' assistance, TA is performed using HTA method. We compile a list of subtasks and concrete operations which are helpful for analyzing the HEs that lead to the failure. HTA includes a set of hierarchical tasks that provide a systematic description of the flight mission of the helicopter. Table 2 shows the subordinate goals and all concrete operations. The subordinate goals i.e. sub-tasks are 1) climb, 2) cross over the bridge, 3) descend, and 4) search and avoid the airbus. Each subordinate goal is further divided into concrete operations.

The brainstorming session with four raters allows us to identify the tasks that should be performed at the same time and their priorities. We then determine the plans of these sub-tasks, through which the work processes are well structured based on its hierarchical approach. Such plans and task priorities is the basis of mental workload analysis.

5.3. Human error analysis using CREAM

Based on the results of TA, CREAM method is introduced to identify and quantify possible HEs. Table 2 shows the detailed operation procedure, and for each operation, we can identify its cognitive function. According to the determined situation, CPC assessment can be conducted with four raters' assistance. For example, Table 3 shows the CPCs for subtask 3. Then weighting factors for CPCs can be determined and the CFP for each operation can be calculated using the extended CREAM method [13].

The methods that combine TA and CREAM for human error identification and quantification have been widely studied in many literatures [34, 54, 55]. Based on such methods, the possible helicopter's errors when performing flight mission can be identified and quantified. In this paper, we focus more on the occurrence and effects of MWLOL, which will be analyzed in detail next.

5.4. Perform a traditional FTA of the accident

The accident report shows that the helicopter crashed during the descent stage (subtask 3.1). The pilot performed subtask 3.1, and task 4 simultaneously at that time. Based on section 5.1-5.3, we gather the HEs, mechanical failures, and environment factors which are combined to cause the accident, and perform a traditional FTA of the helicopter crash accident, as shown in Fig. 7. The helicopter crash during descent stage is due to three categories of causes: 1) helicopter's altitude is too low caused by equipment failures (G1), 2) helicopter's

Table 2. HTA of the flight mission of the helicopter based on [36]

| |
|---|
| 0. Flight mission of the helicopter Plan 0: Do 1 then 2 then 3, and Do 4 simultaneously |
| 1 Climb Plan 1: Do 1.1 then 1.2 then 1.3 |
| 1.1 Determine the climbing position and report the position and climbing request to the controller |
| 1.2 Enter the climb Plan 1.2: Do 1.2.1 then 1.2.2 then 1.2.3 |
| 1.2.1 Increase the collective and throttle, and adjust the pedals as necessary to maintain the longitudinal trim |
| 1.2.2 Move cyclic stick slightly to direct all of the increased power into lift and maintain the airspeed |
| 1.2.3 Check the view and flight instruments to maintain the climb attitude, course, speed, rate of climb, propeller speed, and longitudinal trim until moving to level flight |
| 1.3 Level off the climb Plan 1.3: Do 1.3.1 then 1.3.2 then 1.3.3 then 1.3.4 |
| 1.3.1 Determine the attitude to lead the level-off |
| 1.3.2 Apply forward cyclic stick to adjust the helicopter to level flight attitude |
| 1.3.3 Maintain climb power until the airspeed approaches the desired cruising airspeed, then lower the collective to obtain cruising power and adjust the throttle to obtain and maintain cruising rpm. |
| 1.3.4 Throughout the level-off, control anti-torque pedals to complete longitudinal trim |
| 2 cross over the bridge Plan 2: Do 2.1 then 2.2 then 2.3 then 2.4 then 2.5 |
| 2.1 Apply forward pressure on the cyclic stick forward to obtain the forward speed |
| 2.2 Control the collective pitch lever to maintain the flight attitude |
| 2.3 Control the throttle to maintain the propeller speed |
| 2.4 Control anti-torque pedals to maintain the trim |
| 2.5 Check the view and flight instruments to maintain the climb attitude, course, speed, rate of climb, propeller speed, and trim until moving to descent |
| 3 Descent Plan 3: Do 3.1 then 3.2 |
| 3.1 Enter the decent stage Plan 3.1: Do 3.1.1 then 3.1.2 then 3.1.3 then 3.1.4 then 3.1.5 |
| 3.1.1 Lower collective pitch to obtain proper power |
| 3.1.2 Control the throttle to maintain rpm |
| 3.1.3 Control anti-torque pedals to complete longitudinal trim and maintain the course |
| 3.1.4 Adjust cyclic stick to maintain the descent attitude and speed |
| 3.1.5 Check the view and flight instruments to maintain the power, altitude, course, and longitudinal trim until moving to level flight |
| 3.2 Level off Plan 3.2: Do 3.2.1 then 3.2.2 then 3.2.3 then 3.2.4 |
| 3.2.1 Determine the desired altitude to lead the level-off |
| 3.2.2 Increase collective pitch and throttle to obtain cruising power and maintain rpm |
| 3.2.3 Control anti-torque pedals to complete longitudinal trim and maintain the course |
| 3.2.4 As the helicopter decreases to the required flight altitude, control the cyclic stick to obtain the cruise speed and straight-and-level attitude |
| 4 Search and avoid the airbus Plan 4: Do 4.1 then 4.2 then 4.3 |
| 4.1 Contact air traffic controller for the airbus location |
| 4.2 Search the airbus visually until have the airbus insight |
| 4.3 Control the helicopter and maintain visual separation from the airbus |

altitude is too low caused by extreme environment (G3), and 3) helicopter crashes due to the failure of man-machine interaction (G2). In this accident, the pilot is well trained. If he is aware of the low altitude, he will take measures to prevent helicopter crash. The flight nurse who survived the accident stated: “*the pilot did not execute any evasive maneuvers or communicate any difficulties, either verbally*

or nonverbally” [27]. Therefore, G2 is caused by the pilot perception failure of low altitude.

The MWLLOL occurs during man-machine interaction. In addition, G1 and G3 can be analyzed by traditional FTA methods. Thus we focus more on the cause G2, and it is analyzed by the connection of logic gates and lower events until all its branches are terminated with basic events.

Table 3. Common performance condition assessment for the operations of subtask 3

| CPC name | Level |
|---|---------------|
| Adequacy of organization | Improved |
| Working condition | Reduced |
| Adequacy of man-machine interface and operational support | Insignificant |
| Availability of procedures/plans | Insignificant |
| Number of simultaneous goals | Reduced |
| Available time | Insignificant |
| Time of day (circadian rhythm) | Reduced |
| Adequacy of training and expertise | Improved |
| Crew collaboration quality | Insignificant |

summed to be $1e-4$, the probabilities of HEs (i.e. X2, X4 and X5) are calculated using extended CREAM method.

5.5. Abnormal situation handling tasks caused by basic events

When traditional FT is conducted, the basic events are analyzed first to determine whether they can introduce abnormal situation handling tasks in the process of man-machine interaction. When the basic event X1 “Radar altimeter failure” occurs, the pilot should fly with VFR. In addition, if the basic events X6 “Night flight” and X7 “No outside visual reference” also occur, the pilot cannot obtain altitude from the view. Then a new abnormal situation handling task (i.e. a communication task with ATC for altitude) is introduced. At the same time, the pilot should perform another dual tasks (i.e. subtask 3.1, task 4). Therefore, the new abnormal situation handling task will increase pilot’s mental workload, and may lead to MWLOL. During these concurrent tasks, the MWLOL may lead to the abandonment of communication task with ATC, and then the helicopter crashes into the river due to perception failure of low altitude.

5.6. Mental workload analysis

The pilot performs subtask 3.1, task 4, and communication task simultaneously at that time. We calculate the task interference of these three time-shared tasks based on the extension of MRM, and implement FTA with the MWLOL gate. For these three tasks, priority is given to safe helicopter control (i.e. subtask 3.1 denoted T_1). The secondary task is searching and avoiding the airbus (i.e. task 4 denoted T_2). The communication task is an important but low-priority task because it is not urgent. Thus the communication task denoted T_3 is the third priority task.

Each task is coded by the extent to which it depends on separate resources defined by 4 dimensions mentioned above, as shown in Fig. 1. The pilot performs T_1 following VFR. He views outside and controls the cyclic stick, collective pitch lever, and anti-torque pedals to maintain the rate of decent, propeller speed, course, and longitudinal trim. T_1 can be coded as: Perception: Visual Ambient (=1), Response: Spatial (=2). When performing T_2 , the pilot should do a conversational task with controller and search for the airbus to maintain visual separation. Task 4 can be coded as: Perception: Auditory Verbal (=1), and Visual Ambient (=1). T_3 requires the pilot to ask the ATC for altitude. Such task can be coded as Perception: Auditory Verbal (=1). Thus each task spawns a demand vector: $\mathbf{d}_1 = \{0, 1, 0, 0, 0, 2, 0\}$, $\mathbf{d}_2 = \{0, 1, 0, 1, 0, 0, 0\}$, and $\mathbf{d}_3 = \{1, 0, 0, 0, 0, 0, 0\}$.

Then by querying Table 1, the resource-conflict scores can be obtained. The conflict matrix of T_1T_2 , T_1T_3 , and T_2T_3 is constructed respectively, as shown in Table 5-7.

The resource conflict score of T_1T_2 is equal to the summation of conflict values in Table 5,

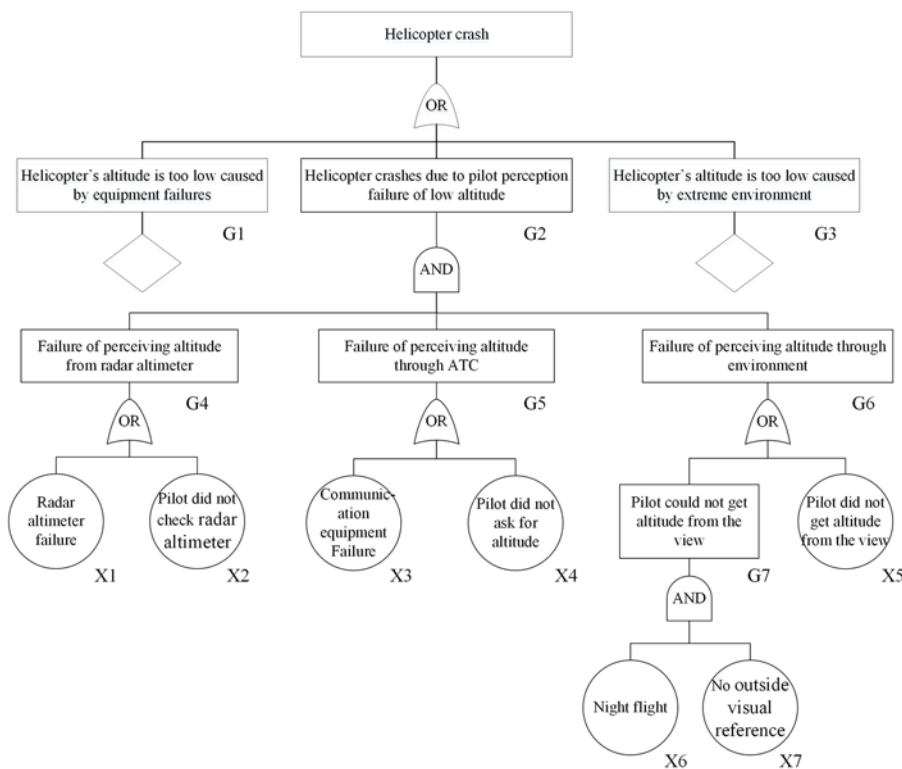


Fig. 7. Traditional FT of the helicopter crash

Based on this traditional FT, the accident of helicopter crash can be analyzed in qualitative and quantitative ways. To analyze quantitatively, the probabilities of the basic events are shown in Table. 4, where the probabilities of equipment failures (i.e. X1 and X3) are as-

Table 4. Probabilities of basic events

| Events | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|-------------|--------|--------|--------|--------|--------|-----|-----|
| Probability | 0.0001 | 0.0269 | 0.0001 | 0.0092 | 0.0269 | 0.5 | 0.1 |

Table 5. Conflict matrix of T_1T_2

| | | T_1 | |
|-------|----|-------|-----|
| | | Va | Rs |
| T_2 | Va | 0.8 | 0.2 |
| | Av | 0.6 | 0.2 |

Table 6. Conflict matrix of T_1T_3

| | | T_1 | |
|-------|----|-------|-----|
| | | Va | Rs |
| T_3 | Av | 0.6 | 0.2 |

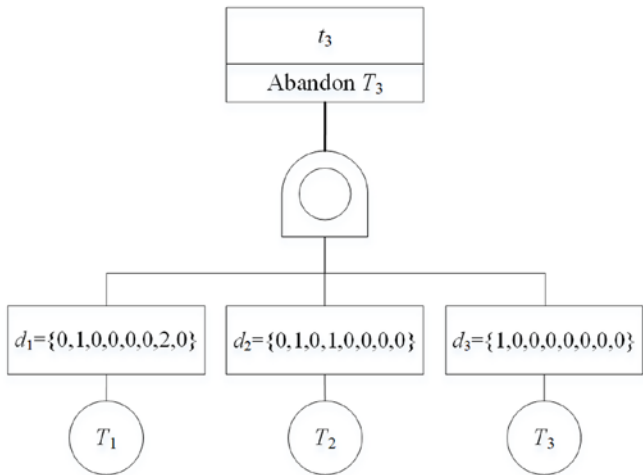
Table 7. Conflict matrix of T_2T_3

| | | T_2 | |
|-------|----|-------|-----|
| | | Va | Av |
| T_3 | Av | 0.6 | 0.8 |

i.e., $R(d_1, d_2) = 0.8 + 0.2 + 0.6 + 0.2 = 1.8$. Then the total interference value of $T_1 T_2$ can be calculated as: $TI_{1+2} = 1 + 2 + 1 + 1 + 1.8 = 6.8$. Based on the extension of MRM, the resource conflict score of $T_1 T_2 T_3$ is equal to the summation of conflict values in Tables 5-7, i.e., $R(d_1, d_2, d_3) = 1.8 + 0.6 + 0.2 + 0.6 + 0.8 = 4.0$. The total interference value of $T_1 T_2 T_3$ can be calculated as: $TI_{1+2+3} = 1 + 1 + 2 + 1 + 1 + 4.0 = 10$.

5.7. MWLOL gate establishing

TI_{1+2+3} exceeds the threshold of 7, i.e., this task scenario actually leads to MWLOL. Performing T_1 , T_2 , and T_3 simultaneously requires a large number of cognitive resources that the pilot is unable to all provide. Based on the mechanism of MWLOL, the pilot will abandon the low-priority tasks until the total interference is less than 7. Thus the pilot will abandon T_3 and the total interference value of $T_1 T_2$ is 6.8. The MWLOL gate can be established. As shown in Fig.8,



t_3 denote the event that abandon T_3 .

Fig. 8. MWLOL Gate of $T_1 T_2 T_3$

5.8. Modelling FTA with MWLOL gate

Performing $T_1 T_2 T_3$ at the same time leads to the MWLOL, and then T_3 will be abandoned. Accordingly, the basic event X4 in Fig 7 i.e. “Pilot did not ask for altitude” is triggered. Therefore, X4 occurs not only due to omission, but also due to MWLOL. Then the event “Pilot did not ask for altitude” denoted as X4 becomes an intermediate event in the modified FTA, denoted as Gn. Gn can be triggered by MWLOL or operation omission, where MWLOL can be described by the MWLOL gate and the omission is basic event whose probability can be calculated by CREAM method. In addition, because the basic event “omission” is the same as the event “Pilot did not ask for altitude” in tradition FT, it is also denoted as X4 in modified FT for the purpose of comparative analysis between modified FTA and traditional FTA.

The traditional FT is modified by the MWLOL gate, and the modified FT is shown in Figure 9. X1, X6, X7, and X4 in traditional FT are independent, while X1, X6, and X7 trigger X4 when considering MWLOL. Through the MWLOL gate, such logic relationship is added explicitly to traditional FT, which is more helpful for analyzing the reasons of HEs and preventing the accident.

6. Results and discussions

6.1. Risk analysis of helicopter crash

As shown in Fig.9, the FT of helicopter crash has been constructed with the MWLOL gate, whose top event is “Helicopter crash due to pilot perception failure of low altitude”. The FT thus created is analyzed through evaluating MCS. The MCS are identified as follows:

$$MCS = \{X1, X3, X5\}, \{X1, X4, X5\}, \{X1, X6, X7\}, \{X2, X3, X5\}, \{X2, X4, X5\}, \{X2, X3, X6, X7\}, \{X2, X4, X6, X7\}$$

Based on the MCS above and the probabilities of basic events shown in Table 4, the probability of top event can be calculated by quantitative methods of traditional FT. In Fig. 9, the helicopter crash probability is $2.392e-5$.

To identify the crucial basic events, we calculate and analyze their probability importance degrees as shown in Table 8. Comparing with the other basic events, X1 (Radar altimeter failure) is the most crucial event. This can be explained by the fact that the combination of X1, X6 (Night flight) and X7 (No outside visual reference) belongs to the MCS and X6 and X7 are high probability events. Therefore, the accident probability is sensitive to X1. X6 and X7 make pilot fly into actual instrument meteorological conditions. X1 combined with X6 and X7 will introduce pilot’s communication task with ATC for altitude. Then the MWLOL lead to pilot’s perception failure of low altitude and finally the helicopter crashes.

In addition, X3 (Communication equipment failure) and X4 (Omission) shall also attract more attention because these two events are relatively more crucial than the others apart from X1.

6.2. Comparison with the traditional FTA

As shown in Fig.7, the traditional FT of helicopter crash has been constructed. The MCS are identified as follows:

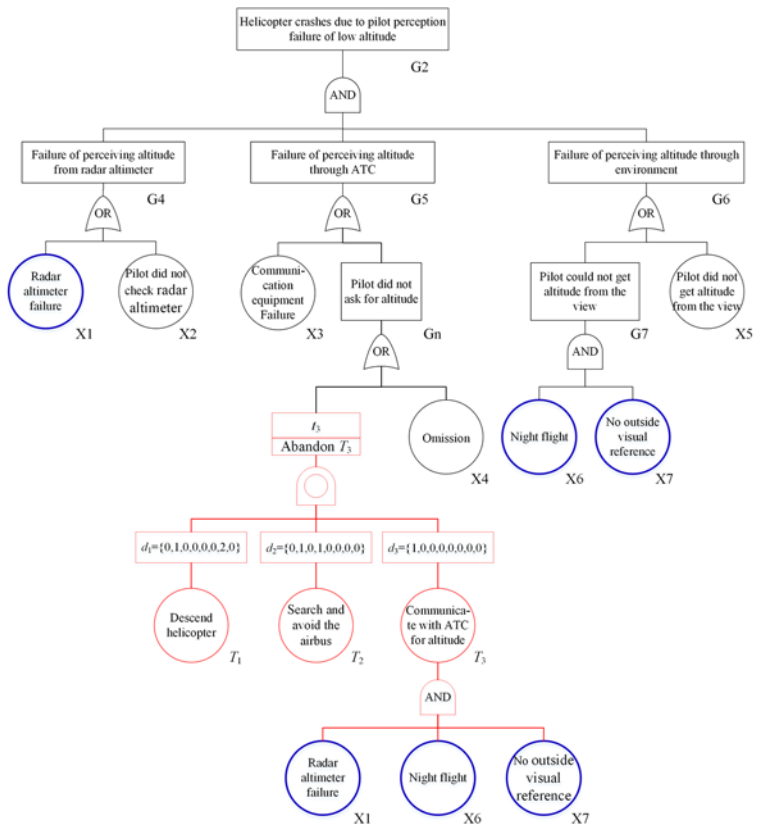


Fig. 9 Modified FT of the helicopter crash accident

Table 8. Probability importance degrees of basic events in traditional FT

| Basic Event | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|------------------------|----------|----------|----------|----------|----------|----------|----------|
| Probability importance | 5.02E-02 | 7.03E-04 | 2.02E-03 | 2.03E-03 | 2.38E-04 | 3.43E-05 | 1.72E-04 |

Table 9. Probability importance degrees of basic events in the modified FT

| Basic Event | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|------------------------|----------|----------|----------|----------|----------|----------|----------|
| Probability importance | 6.84E-04 | 7.03E-04 | 2.02E-03 | 2.03E-03 | 2.38E-04 | 2.44E-05 | 1.22E-04 |

$$\text{MCS} = \{X1, X3, X5\}, \{X1, X3, X6, X7\}, \{X1, X4, X5\}, \{X1, X4, X6, X7\}, \\ \{X2, X3, X5\}, \{X2, X3, X6, X7\}, \{X2, X4, X5\}, \{X2, X4, X6, X7\}$$

Without MWLOL gate, the logic relationship among X1, X4 (Pilot did not ask for altitude), X6 and X7 (i.e. X1, X6, and X7 trigger X4) cannot be described. X1 combined with X6 and X7 cannot lead to the occurrence of top event. MCS $\{X1, X6, X7\}$ in modified FT is changed to MCS $\{X1, X3, X6, X7\}$, and $\{X1, X4, X6, X7\}$ in traditional FT. Based on this MCS, the helicopter crash probability is $1.897\text{e-}5$ which is 26.1% lower than the modified FTA method. In addition, the importance degrees of X1 decreases significantly as shown in Table 9.

6.3. Analysis of MWLOL's contribution to helicopter crash

As discussed above, the combination of X1, X6, and X7 will introduce T_3 (pilot's communication task with ATC for altitude). T_1, T_2 , and T_3 are time-shared, which lead to MWLOL. Then T_3 is abandoned, and pilot cannot be aware of helicopter altitude information. Finally, helicopter crashes during descent stage due to pilot perception failure. The above man-machine interaction process is described through MWLOL gate. The logic relationship that X1, X6, and X7 trigger X4 is established and $\{X1, X6, X7\}$ belongs to MCS. Therefore, when considering MWLOL the helicopter crash is more likely to occur and helicopter crash probability increases from $1.897\text{e-}5$ to $2.392\text{e-}5$. In addition, X1 becomes more crucial, its importance degree increases from $6.84\text{E-}04$ to $5.02\text{E-}02$. If the helicopter flies with an inoperative radar altimeter, the top event probability in the modified FT with MWLOL gate is 0.0502. In addition, if this helicopter flies at night, the top event probability is 0.1002. Therefore, through MWLOL gate, the analysis shows that X1 is a weakness of the helicopter system especially flying at night.

If ignoring MWLOL, the importance degree of X1 is $6.84\text{E-}04$, and the most crucial basic events will be regarded as X3 and X4. In addition, if the helicopter flies with an inoperative radar altimeter, the top event probability in traditional FT without MWLOL gate is $7.02\text{e-}4$. Comparing with the results of modified FTA (i.e. 0.0502), such accident probability decreases significantly. The traditional FT cannot identify the true root causes of accident. Accordingly, it is impossible to take targeted measures to prevent accidents.

6.4. Validation and suggestions

For this accident, the NTSB determines that "the probable cause of this accident was the pilot's failure to identify and arrest the helicopter's descent, which resulted in controlled flight into terrain. Contrib-

uting to the accident were the dark night conditions, limited outside visual references, and the lack of an operable radar altimeter in the helicopter." [27]. That is to say, the occurrence of X1, X6, and X7 leads to the pilot's perception failure and then results in helicopter crash. Such accident causes demonstrate the need to consider MWLOL and the effectiveness of the modified FTA

Based on the results of modified FTA and the accident causes determined by NTSB, radar altimeter is a vulnerability of the helicopter system when flying at night because it is necessary to ensure altitude awareness when the helicopter

flies into instrument meteorological conditions. Therefore, the radar altimeter should be pay more attention when performing aircraft inspection program. However, the radar altimeter is out of the FAA-approved Minimum Equipment List (MEL) and can be deferred for maintenance within 10 calendar days. In this accident, the maintenance logbook on January 10, 2005 included an entry for an inoperative radar altimeter. According to "MEL Items and Deferred Maintenance" section, the inoperative radar altimeter could be deferred for maintenance until January 20, 2005. Then the helicopter with an inoperative radar altimeter was allowed to perform flying tasks, and the inoperative radar altimeter lead to this accident. Therefore, when flying at night, the radar altimeter should be added to the MEL.

7. Conclusions

Effective risk analysis and accident prevention need analyze pilot mental workload to better understand human behavior in accident occurrence. In this paper, a MRM is introduced to analyze mental workload in risk analysis, and a MWLOL gate is first proposed to incorporate MWLOL into previous FTA methods combined with TA and HRA. The proposed risk analysis method modifies traditional FTA through the MWLOL gate, while it retains the analytical capability of traditional FTA. It provides a more in-depth risk analysis of man-machine system, and it can also assess the technical safety of machine system. In addition, the proposed method models the normal task and abnormal situation handling task as a whole, and analyzes all possible events to assess the risk of systems. Therefore, the risk analysis may be more comprehensive.

This modified FTA is successfully used to analyze accident for the first time. As seen from the case study, through the MWLOL gate, logic relationships among basic events due to the MWLOL in the process of handling abnormal situations are added to traditional FT. Comparing with the results of traditional FTA, the modified FTA obtains more rational MCS, important degrees of basic events, and top event probability, which are validated by a case study of helicopter crash in Maryland reported by NTSB. Last but not least, an insight of the causes of the helicopter crash accident in Maryland is gained and some suggests are given to prevent future similar accidents.

References

1. Abilio Ramos M, Utne I B, Mosleh A. Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. Safety Science 2019;116:33-44, <http://doi.org/10.1016/j.ssci.2019.02.038>.
2. Blom H, Daams J, Nijhuis H. Human cognition modelling in ATM safety assessment: Nat Aero Lab 2001. NLR, NLR-TP-2001-053, 2001,

- <https://www.researchgate.net/publication/228817723>.
3. Bolton M L, Molinaro K A, Houser A M. A formal method for assessing the impact of task-based erroneous human behavior on system safety. *Reliability Engineering & System Safety* 2019;188:168-180, <http://doi.org/10.1016/j.ress.2019.03.010>.
 4. Cline P. Human Error Analysis of Helicopter Emergency Medical Services (HEMS) Accidents Using the Human Factors Analysis and Classification System (HFACS). *Journal of aviation/aerospace education and research* 2018;1(28):43-62, <http://doi.org/10.15394/jaaer.2018.1758>.
 5. DiDomenico A, Nussbaum M A. Effects of different physical workload parameters on mental workload and performance. *International Journal of Industrial Ergonomics* 2011;41(3):255-260, <http://doi.org/10.1016/j.ergon.2011.01.008>.
 6. Doytchev D E, Szwillus G. Combining task analysis and fault tree analysis for accident and incident analysis: A case study from Bulgaria. *Accident Analysis & Prevention* 2009;41(6):1172-1179, <http://doi.org/10.1016/j.aap.2008.07.014>.
 7. Dugan J B, Bavuso S J, Boyd M A. Dynamic fault-tree models for fault-tolerant computer systems. *IEEE Transactions On Reliability* 1992;41(3):363-377, <http://doi.org/10.1109/24.159800>.
 8. García-Mas A, Ortega E, Ponseti J, De Teresa C, Cárdenas D. Workload and cortisol levels in helicopter combat pilots during simulated flights. *Revista Andaluza de Medicina del Deporte* 2015;9(1):7-11, <http://doi.org/10.1016/j.ramd.2015.12.001>.
 9. Gore B F, Jarvis P. Modeling the complexities of human performance. *IEEE International Conference on Systems* 2006, <http://doi.org/10.1109/ICSMC.2005.1571377>.
 10. Gregoriades A, Sutcliffe A. Workload prediction for improved design and reliability of complex systems. *Reliability Engineering & System Safety* 2008;93(4):530-549, <http://doi.org/10.1016/j.ress.2007.02.001>.
 11. Hankins T C, Wilson G F. A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space & Environmental Medicine* 1998;69(4):360, <http://doi.org/10.1111/j.1445-5994.1998.tb02982.x>.
 12. Hollnagel E. The phenotype of erroneous actions. *International Journal of Man-Machine Studies* 1993;1(39):1-32, <https://www.researchgate.net/publication/239654361>
 13. Hollnagel E. *Cognitive reliability and error analysis method (CREAM)*. Oxford, UK: Elsevier, 1998, <https://doi.org/10.1016/B978-008042848-2/50010-5>.
 14. Kabir S. An overview of fault tree analysis and its application in model based dependability analysis. *Expert Systems with Applications* 2017;77:114-135, <http://doi.org/10.1016/j.eswa.2017.01.058>.
 15. Khakzad N, Khan F, Amyotte P. Risk-based design of process systems using discrete-time Bayesian networks. *Reliability Engineering & System Safety* 2013;109:5-17, <http://doi.org/10.1016/j.ress.2012.07.009>.
 16. Khakzad N, Khan F, Amyotte P. Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. *Reliability Engineering & System Safety* 2011;96(8):925-932, <http://doi.org/10.1016/j.ress.2011.03.012>.
 17. Kirwan B, Ainsworth L K. *A Guide to Task Analysis*. London: Taylor & Francis, 1992, <https://doi.org/10.1201/b16826>.
 18. Liu C, Kramer A, Neumann S. Reliability assessment of repairable phased-mission system by Monte Carlo simulation based on modular sequence-enforcing fault tree model. *Eksploatacja i Niezawodność - Maintenance and Reliability* 2020;22(2):272-281, <http://doi.org/10.17531/ein.2020.2.10>.
 19. McCracken J, Aldrich T. *Analyses of Selected LHX Mission Functions: Implications for Operator Workload and System Automation Goals*. Defense Technical Information Center 1984.
 20. Mohaghegh Z, Mosleh A. Incorporating organizational factors into probabilistic risk assessment of complex socio-technical systems: Principles and theoretical foundations. *Safety Science* 2009;47(8):1139-1158, <http://doi.org/10.1016/j.ssci.2008.12.008>.
 21. Naderpour M, Lu J, Zhang G. An abnormal situation modeling method to assist operators in safety-critical systems. *Reliability Engineering & System Safety* 2015;133:33-47, <http://doi.org/10.1016/j.ress.2014.08.003>.
 22. Nees M A, Sharma N, Shore A. Attributions of accidents to "human error" in news stories: Effects on perceived culpability, perceived preventability, and perceived need for punishment. *Accident Analysis & Prevention* 2020;148:105792, <http://doi.org/10.1016/j.aap.2020.105792>.
 23. Noel J B, Bauer K W, Lanning J W. Improving pilot mental workload classification through feature exploitation and combination: a feasibility study. *Computers & Operations Research* 2005;32(10):2713-2730, <http://doi.org/10.1016/j.cor.2004.03.022>.
 24. Papadimitriou E, Schneider C, Aguinaga Tello J, Damen W, Lomba Vrouwenraets M, Ten Broeke A. Transport safety and human factors in the era of automation: What can transport modes learn from each other? *Accident Analysis & Prevention* 2020;144:105656, <http://doi.org/10.1016/j.aap.2020.105656>.
 25. Papis M, Jastrzębski D, Kopyt A, Matyjewski M, Mirosław M. Driver reliability and behavior study based on a car simulator station tests in ACC system scenarios. *Eksploatacja i Niezawodność - Maintenance and Reliability* 2019;21(3):511-521, <http://doi.org/10.17531/ein.2019.3.18>.
 26. Qingxin Z, Zhuang D, Yinxian M. Design of target code in human-machine interface. *Journal of Beijing University of Aeronautics and Astronautics* 2007;33(06):631-634, <http://doi.org/10.13700/j.bh.1001-5965.2007.06.001>.
 27. Rosenker M V, Sumwalt R L, Hersman D A P, Higgins K O, Chealander S R. *Aviation Accident Report: NYC05MA039*. Washington: National Transportation Safety Board, 2007.
 28. Ruijters E, Stoelinga M. Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. *Computer Science Review* 2015;15-16:29-62, <http://doi.org/10.1016/j.cosrev.2015.03.001>.
 29. Schulte A, Donath D, Honecker F. Human-System Interaction Analysis for Military Pilot Activity and Mental Workload Determination.: *IEEE*, 2015. 1375-1380, <http://doi.org/10.1109/SMC.2015.244>.
 30. Shappell S A, Wiegmann D A. *HFACS Analysis of Military and Civilian Aviation Accidents: A North American Comparison*. ISASI 2004:1-8.
 31. Sheridan T B, Parasuraman R. Human-Automation Interaction. *Reviews of Human Factors and Ergonomics* 2005;1(1):89-129, <https://www.researchgate.net/publication/240756917>.
 32. Stanton N A. Hierarchical task analysis: Developments, applications, and extensions. *Applied Ergonomics* 2006;37(1):55-79, <http://doi.org/10.1016/j.apergo.2005.06.003>.
 33. Sussman D, Coplen M. Fatigue and alertness in the United States railroad industry part I: The nature of the problem. *Transportation Research Part F Traffic Psychology and Behaviour* 2000;3(4):211-220, [http://doi.org/10.1016/S1369-8478\(01\)00005-5](http://doi.org/10.1016/S1369-8478(01)00005-5)

34. Ung S. Human error assessment of oil tanker grounding. *Safety Science* 2018;104:16-28, <http://doi.org/10.1016/j.ssci.2017.12.035>.
35. Gawron V.J. Summary of Fatigue Research for Civilian and Military Pilots. *IIE Transactions on Occupational Ergonomics and Human Factors* 2015;4(1):1-18, <https://doi.org/10.1080/21577323.2015.1046093>
36. Veillette P. Helicopter Flying Handbook. U.S. Department of Transportation United States Department of Transportation, Federal Aviation Administration, Airman Testing Standards Branch, AFS-630, P.O. Box 25082, Oklahoma City, OK 73125: Flight Standards Service, 2012.
37. Walker M, Bottaci L, Papadopoulos Y. Compositional Temporal Fault Tree Analysis. *Computer Safety, Reliability, & Security, International Conference, Safecomp, Nuremberg, Germany, September 2007*, https://link.springer.com/chapter/10.1007%2F978-3-540-75101-4_12.
38. Wang P, Fang W, Guo B. A colored petri nets based workload evaluation model and its validation through Multi-Attribute Task Battery-II. *Applied Ergonomics* 2017;60:260-274, <http://doi.org/10.1016/j.apergo.2016.11.013>.
39. Wang W, Jiang X, Xia S, Cao Q. Incident tree model and incident tree analysis method for quantified risk assessment: An in-depth accident study in traffic operation. *Safety Science* 2010;48(10):1248-1262, <http://doi.org/10.1016/j.ssci.2010.04.002>.
40. Wickens C D. Multiple Resources and Mental Workload. *Human Factors*; 50(3):449-455, <http://doi.org/10.1518/001872008X288394>.
41. Wickens C D. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* 2002;2(3):159-177, <http://doi.org/10.1080/14639220210123806>
42. Wickens C D, Goh J, Helleberg J, Horrey W J, Talleur D A. Attentional models of multitask pilot performance using advanced display technology. *Human Factors* 2003;45(3):360-380, <http://doi.org/10.1518/hfes.45.3.360.27250>
43. Wickens C D, Hollands J G, Simon B, Parasuraman R. *Engineering Psychology and Human Performance*. USA: Pearson Education, Inc., 2012.
44. Wickens C, Colcombe A. Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors* 2007;49(5):839-850, <http://doi.org/10.1518/001872007X230217>.
45. Wijayarathna P G, Maekawa M. Extending fault trees with an AND-THEN gate.: *IEEE*, 2000. 283-292, <http://doi.org/10.1109/ISSRE.2000.885879>
46. Wu J, Wen H, Qi W. A new method of temporal and spatial risk estimation for lane change considering conventional recognition defects. *Accident Analysis & Prevention* 2020;148:105796, <http://doi.org/10.1016/j.aap.2020.105796>.
47. Xu Z, Guo D, Wang J, Li X, Ge D. A numerical simulation method for a repairable dynamic fault tree. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2021;23(1):34-41, <http://doi.org/10.17531/ein.2021.1.4>.
48. Young M S, Brookhuis K A, Wickens C D, Hancock P A. State of science: mental workload in ergonomics. *Ergonomics* 2015;58(1):1-17, <http://doi.org/10.1080/00140139.2014.956151>.
49. Zarei E, Yazdi M, Abbassi R, Khan F. A hybrid model for human factor analysis in process accidents: FBN-HFACS. *Journal of Loss Prevention in the Process Industries* 2019;57:142-155, <http://doi.org/10.1016/j.jlp.2018.11.015>.
50. Zhang H, Zhuang D, Ma D, Sun J. Design and Evaluation of Target Icons of the Cockpit Display Interface. *Packaging Engineering* 2011;32(10):89-92, <http://doi.org/10.3354/cr00999>.
51. Zhang X, Qu X, Xue H, Zhao H, Li T, Tao D. Modeling pilot mental workload using information theory. *The Aeronautical Journal* 2019;123(1264):828-839, <http://doi.org/10.1017/aer.2019.13>.
52. Zhang Y, Zheng H, Duan Y, Meng L, Zhang L. An integrated approach to subjective measuring commercial aviation pilot workload.: *IEEE*, 2015. 1098-1103. <http://doi.org/10.1109/ICIEA.2015.7334270>.
53. Zhou J, Lei Y, Chen Y. A hybrid HEART method to estimate human error probabilities in locomotive driving process. *Reliability Engineering & System Safety* 2019;188:80-89, <http://doi.org/10.1016/j.ress.2019.03.001>.
54. Zhou Q, Wong Y D, Loh H S, Yuen K F. A fuzzy and Bayesian network CREAM model for human reliability analysis – The case of tanker shipping. *Safety Science* 2018;105:149-157, <http://doi.org/10.1016/j.ssci.2018.02.011>.
55. Zhou T, Wu C, Zhang J, Zhang D. Incorporating CREAM and MCS into fault tree analysis of LNG carrier spill accidents. *Safety Science* 2017;96:183-191, <http://doi.org/10.1016/j.ssci.2017.03.015>.

Evaluation of efficiency and reliability of airport processes using simulation tools

Paweł Gołda^a, Tomasz Zawisza^b, Mariusz Izdebski^{c*}

Air Force Institute of Technology, IT Logistics Support Division, ul. Księcia Bolesława 6, Warsaw, Poland

National Cyber Security Centre, ul. Rakowiecka 2, Warsaw, Poland

Warsaw University of Technology Faculty of Transport, 75 Koszykowa str. Warsaw, Poland

Indexed by:




Highlights

- The genetic algorithm to evaluate the effectiveness of take-offs and landings is shown.
- The model for assessing the implementation of airport processes is developed.
- The proposed simulation tool reduces the landing time of aircraft.
- The optimization of taxi routes affects the reliability of airport processes.

Abstract

The purpose of this paper is to evaluate the efficiency of airport processes using simulation tools. A critical review of selected scientific studies relating to the performance of airport processes with respect to reliability, particularly within the apron, has been undertaken. The developed decision-making model evaluates the efficiency of airport processes in terms of minimizing penalties associated with aircraft landing before or after the scheduled landing time. The model takes into account, among other things, aircraft take-offs and landings and separation times between successive aircraft. In order to be able to verify the correctness of the decision-making model, a simulation tool was developed to support decision making in the implementation of airport operations based on a genetic algorithm. A novel development of the structure of a genetic algorithm as well as crossover and mutation operators adapted to the determination of aircraft movement routes on the apron is presented. The developed simulation tool was verified on real input data.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

efficiency, airport processes, simulation tools, safety of airport operations, genetic algorithm, reliability of airport processes.

1. Introduction

The concept of transport or logistics process efficiency is widely discussed in the literature and is interpreted differently depending on the analysed research problem, e.g., efficiency of supply chains [20], production processes [8], intermodal transport [25], railway transport [46], international transport [22], or efficiency of means of transport studied in the context of minimizing exhaust emissions [6].

Processes implemented by airport systems are mainly the so-called airside operations, i.e. operations performed near the airport and on its manoeuvring area. These include aircraft take-off, landing and taxiing operations [30], and ground handling [49]. Airport processes are logistics processes that focus on the operations associated with the flow of a passenger stream at a given airport. The efficiency of any logistic, transport process is based on its reliability in carrying out given logistic operations [48]. Reliability of airport processes implementation is considered in the context of efficient functioning of the airport and its ability to serve passengers [42].

The procedures for take-off and landing are an important aspect in the implementation of airport operations. These procedures include several stages (Fig. 1):

- stage 1 – during which the aircraft captain requests permission to taxi for take-off. He receives information about the runway in use and permission to taxi;
- stage 2 – the aircraft taxis along the taxiways to a designated place in front of the runway (if the air traffic situation requires so, the departing aircraft will be stopped at a place safe for the performance of other airport operations);
- stage 3 – in the absence of contraindications to the take-off operation, a take-off permission is issued, if the situation did not allow the issue of such permission in Stage 2;
- stage 4 – a landing permission is issued if there are no factors preventing the landing operation;
- stage 5 – at this stage permission for the aircraft to taxi on the apron is issued;
- stage 6 – information is given on the location of the aircraft's parking on the apron.

At larger airports, the aircraft, after taxiing to a parking area designated by the air traffic coordinator, is connected to the passenger terminal by a mobile jetway. Many researchers [32] identify aircraft taxiing operations on the airport apron as the most important element affecting airport safety, reliability and capacity. In most cases, the

(*) Corresponding author.

E-mail addresses: P. Gołda - pawel.golda@itwl.pl, T. Zawisza - tzawisza@mon.gov.pl, M. Izdebski - mariusz.izdebski@pw.edu.pl

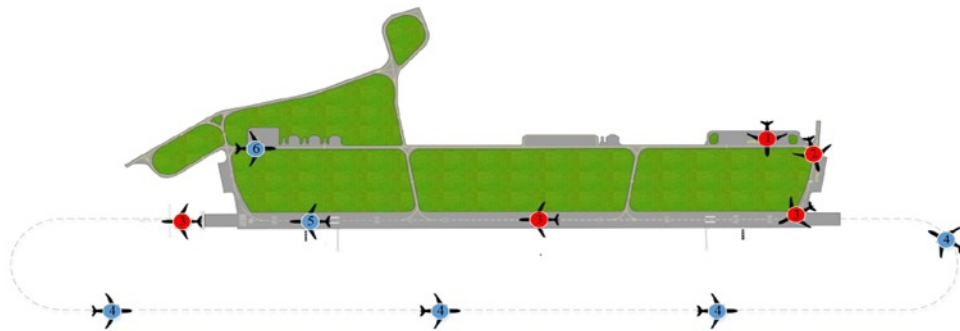


Fig. 1. Specific aircraft positions as seen from the aerodrome control tower

Source: own elaboration based on the developed application for simulation and management of aircraft traffic within the airport.

problem of determining a taxiway is solved solely by taking into account the shortest taxiway distance, disregarding the number of stops and accelerations or the necessity to wait for a free parking space.

Scheduling of take-off and landing operations has received quite a bit of attention in the literature. For example, in the study [45], the authors present a stochastic approach to scheduling take-off operations by describing delays, taxi time, or deviation from the desired arrival time as random variables. The authors emphasize that the aircraft taxiing system is a key factor generating delays in the landing and take-off phases during peak hours. However, scheduling of take-off operations in terms of minimizing potential delays and maximizing airport capacity is also the subject of the paper [35]. The authors propose a dynamic programming based, real-time method to generate a set of potential flight sequences given criteria related to airport delays and capacity. The constraints considered are distance separation, potential taxiway intersections, and separation due to aircraft induced air and exhaust turbulence.

Importantly, many researchers point to the need for a new approach to scheduling and routing of taxiing operations due to the need to maintain adequate safe take-off intervals [40]. In the work presented, the authors proposed an approach using a combinatorial integer optimization task that takes into account the time windows of aircraft entry into an airport's network of ground roads, taxiing speeds, and aircraft stopping characteristics on the apron. On the other hand, in the paper [34], the authors propose an airport taxiway network condition monitoring algorithm using advanced stochastic hybrid linear algorithms.

The main processes determining the reliability and capacity of airports and thus their efficiency are operations of take-off and landing on the runway [4] allocation of gates and parking places [9] and movement of aircraft on the apron [47].

Taking into account the fact that airport processes taking place on the apron affect the reliable and efficient functioning of the airport and determine the safety of passengers, it is advisable to develop modern methods and algorithms to improve safety and minimize the risk of accidents. The authors of this paper presented an original approach to evaluate the efficiency of airport processes by the application of a simulation tool based on a genetic algorithm.

In the first part, a critical analysis of the literature in the described research area is made. Then the author's decision-making model is presented, which includes all the important elements of the process of aircraft management on the airport apron. The model takes into account, among other things, aircraft take-offs and landings and separation times between successive aircraft. The developed decision-making model evaluates the efficiency of airport processes in terms of minimizing penalties associated with aircraft landing before or after the scheduled landing time. The factor that determines the amount of penalties associated with landing an aircraft outside of the designated time windows is the aircraft's taxiing time on the apron. This time will be optimized by the developed simulation tool.

An important element of the article is the verification of the decision-making model and the evaluation of the efficiency of the implementation of airport processes using a simulation tool. The optimization processes in the simulation tool used are implemented by a genetic algorithm. Genetic algorithms are algorithms often used in complex optimization issues, e.g. vehicle routing issue [24] in supply chain design [23] in airspace traffic management [12].

2. Research problems of airport process management - analysis of the literature

2.1. Decision-making problems in air traffic management on the airport apron

The movement of aircraft on the apron is actually a set of scheduling problems and finding the most advantageous route. It is about transit of the aircraft on the ground routes at the airport in such a way that they can achieve their objectives within a given time, i.e.: to reduce the overall travel time and to match the arrival and departure time windows of other aircraft using the airport, bearing in mind the reliability and safety of all operations.

The issues of finding the most advantageous route show a significant level of complexity, depending on the size of the airport and its traffic load. In simple cases where only a few aircraft are simultaneously moving through an area, there is little risk of collisions occurring. In such cases, well-known algorithms for finding shortest paths in a graph, such as Dijkstra's or A* algorithm, are used. More advanced systems require the use of simulation methods and complex optimization algorithms e.g. ant colony optimization (ACO) algorithms [13].

The aircraft taxiing problem is a complex decision-making issue. The following groups of aircraft taxiing restrictions are encountered in reference literature [43]:

- Maintaining an established taxiway. If a taxiway is designated for non-planning reasons, only the issue of take-off and landing scheduling operations that are preceded by taxiing operations is considered [40]. Another approach is presented in [12] in which the problem solving algorithm selects a taxiway from a set of predefined solutions.
- Separation between aircraft [14]. For the sake of reliability and safety of all airport operations, the need for adequate time and distance intervals between aircraft results from the possibility of a direct collision between them.
- The speed at which aircraft move on the apron. In literature there are various approaches to the problem of determining the taxiing speed. Generally speaking, speed depends on the type of aircraft and the shape of the taxiway (curve characteristics) on which the aircraft is moving.
- Taxiing time restrictions for arriving and departing aircraft. For landing operations, it is assumed that the taxiing time from the

runway to the parking area may be constant or variable within a certain range. In most cases it is assumed that the aircraft taxis to a vacant parking area and is expected to reach it in the shortest possible time. In the case of take-off operations, the matter is more complex as it is necessary to consider the problem of selecting the optimal route and, in addition, the problem of take-off sequencing [30].

The achievable number of aircraft landings and take-offs under certain infrastructure conditions is essential information for planning the expansion of airports with new taxiways, runways and aircraft parking areas. Accurate information can significantly affect financial planning for airport expansion. In addition, accurate information on taxiing times is essential for planning airport operations and thus ensuring their reliability and safety. Air traffic controllers instruct pilots on departures and approaches to parking areas and designated take-off routes [29]. Reliable and predictable taxiing time information takes some of the air traffic coordination burden off the air traffic controller.

The airport ground traffic problem involves planning aircraft movements between airport facilities so as to eliminate traffic conflicts in the most technically, economically, environmentally, and safety efficient manner possible [14]. Thus, it affects the reliability of airport operations.

Each arriving aircraft is directed off the runway to a parking area on the apron, or service area. The departing aircraft must be diverted from its current parking position to the runway. Taxiways for departing aircraft moving from established gates and parking areas to runways are predetermined and if there is a conflict with another aircraft, one aircraft must stop and wait. This situation results in delayed departures and potential delays in reaching the destination or increased travel cost due to the need to increase speed [1].

2.2. Issues of aircraft taxiing on the apron in terms of congestion consequences of aircraft traffic

The limited capacity of the airport associated with the organization of ground traffic results in long waiting times for aircraft to take off. The airport ground traffic problem involves planning aircraft movements between airport facilities so as to eliminate traffic conflicts in the most technically, economically, environmentally, and safety efficient manner possible.

One of the primary indicators for evaluating the quality of work in aircraft handling systems is the punctuality of flight completion. The European Organization for the Safety of Air Navigation points out that the main factors determining flight punctuality are delays due to airport operations, including limited runway access. Minimizing take-off times improves runway safety, ensures good utilization of its capacity and ensures reliability of all operations. Minimizing parking waiting times reduces passenger waiting times, which increases the quality of service.

Taxiing time is the time when the aircraft uses its engines while remaining on the ground. For departures, it is the time between leaving the parking position and take-off; for arrivals, it is the time between landing and reaching the parking position. This includes any waiting time, as well as queuing time, not just time in motion. The primary objective of the research work in this area is to minimize average departure and arrival delay times and average taxi waiting times and the associated safety and environmental impact criteria. Minimization of taxiing time implies reduction of pollutant emissions. The taxiing issue may be broken down into the following elements [2]:

- decisions concerning the aircraft movement path on the apron, to and from the parking position (if not already taken),
- allocation of gates and aircraft parking areas,
- landing (and take-off) sequence decisions where ground routes are already established.

Decision support is most often carried out by developing optimization and simulation models. The importance of the ground traffic

optimization problem is highlighted in [4]. Most of the proposed approaches to solving taxiway determination problems are based on simplified decision-making models based on basic ground traffic information [7].

Most of the available research work is devoted to the analysis of runway access planning using heuristic techniques: genetic and ant colony algorithms [28], or cellular automata [36].

An issue related to taxiing is congestion and its impact on the efficiency of airport operations. This paper [29] presents a model of aircraft taxiing on the apron and two strategies for solving it: varying aircraft departure and arrival times and varying departure times only, which greatly facilitates the use of the model.

In airport processes, the flight controller managing aircraft traffic has access to information on all aircraft and their location in the airspace. In this respect, ground air traffic control is similar to the systems used in Automated Guided Vehicles (AGVs) [10], which are computer controlled.

The problem of aircraft taxiing is widely described in publications [39]. These publications offer some detailed solutions, but do not present a coherent model or methodology for studying and making decisions about the processes of taxiing and handling aircraft at airports and their impact on the efficiency of airport operations.

The literature review has highlighted that it is reasonable to develop new tools to support decision-making in the implementation of airport operations to eliminate conflict situations while minimizing the duration of airport operations [15], which consequently affects the efficiency of all operations.

3. Model of airport process implementation

3.1. Take-off and landing model parameters

The data necessary for the development of a mathematical model for scheduling aircraft take-offs and landings, taking into account the separation times between successive aircraft, the possibility of landing on different runways/landing fields, and the costs of penalties for landing outside the time set are presented below in Table 1.

Table 1. Decision-making model parameters

| Parameter | Description |
|-----------|--|
| I | the set of flight/aircraft numbers, where i, j are elements of the set |
| SL | the set of runways/landing fields, where sl, sl' are elements of the set |
| A_i | the earliest possible time for landing by i -th flight/aircraft |
| B_i | the latest possible time for landing by i -th flight/plane |
| ML_i | planned time of landing by the i -th flight/aircraft |
| kA_i | unit amount of penalty for landing the aircraft before its scheduled time of arrival |
| kB_i | the unit amount of the penalty for landing the aircraft after its scheduled time of arrival |
| TS_{ij} | separation time between the landing of aircraft no. i and aircraft no. j |
| ts_{ij} | separation time between landing of aircraft no. i and aircraft no. j on different runways/landing fields |

3.2. Quantities sought

The decision variables sought in the model relate to the values of aircraft landing times, landing sequence and runways/landing fields. Therefore, the aircraft landing sequence in the model was written in the form of a binary variable (taking the values 1 and 0). On the other

hand, the aircraft landing times were recorded in the form of variables taking values from the set of positive real numbers. The defined decision variables are shown in Tab. 2.

Table 2. The variables sought in the decision model

| Variable | Description |
|------------|--|
| f_{ij} | $f_{ij}=1$ if the i -th aircraft lands before the j -th aircraft; otherwise it takes the value 0; |
| g_{ij} | $g_{ij}=1$ if the i -th aircraft lands on the same runway/landing field as the aircraft no. j ; otherwise it takes the value 0 |
| u_i^{sl} | $u_i^{sl}=1$ if the i -th aircraft lands on sl -th runway/landing field; otherwise it takes the value 0 |
| lm_i | landing time of the i -th aircraft |

3.3. Criterion function and constraints

The criterion function has the interpretation of minimizing penalties associated with landing the aircraft before or after the scheduled landing time:

$$\sum_{i=1} (kA_i (ML_i - lm_i) + kB_i (lm_i - ML_i)) \rightarrow \min \quad (1)$$

The constraints imposed on the values of the decision variables are as follows:

- Each landing must be made within the time interval determined by the earliest and latest landing times:

$$\forall i \in I \quad A_i \leq lm_i \leq B_i \quad (2)$$

$$lm_i = A_i + g_{ij} (B_i - A_i) \quad (3)$$

- Constraint of the sequence in which aircraft land:

$$\forall i, j \in I \quad \wedge \quad j > i \quad f_{ij} + f_{ji} = 1 \quad (4)$$

$$\forall i, j \in I \quad g_{ij} = g_{ji} \quad (5)$$

- Constraint of the separation time between successive landing aircraft:

$$\forall i, j \in I \quad lm_j \geq x_i + TS_{ij}g_{ij} + ts_{ij}(1 - g_{ij}) - M \cdot f_{ij} \quad (6)$$

where: M – is a large number ensuring that this constraint is redundant when aircraft number j lands before aircraft number i .

- Each aircraft is assigned to only one runway/landing field:

$$\forall i \in I \quad \sum_{sl=1} u_i^{sl} = 1 \quad (7)$$

$$\forall i, j \in I \quad \forall sl \in SL \quad g_{ij} \geq u_i^{sl} + u_j^{sl} - 1 \quad (8)$$

4. Application of genetic algorithm in the organization of aircraft traffic on the apron

4.1. General assumptions

The simulation tool developed in this paper to evaluate the efficiency of airport processes is based on the genetic algorithm. The

task of the algorithm is to determine the transit routes of aircraft when they take off and land, taking into account the sequence of their take-offs and landings. These routes will generate apron occupancy times and thus determine the amount of penalties associated with aircraft landing before or after the scheduled landing or take-off time. In addition, the landing times for individual aircraft at the airport are determined based on the apron occupancy times.

The principle of the genetic algorithm can be presented in the following steps:

- Step 1.** Input data introduction: average transit time between point elements of the apron structure, times for additional aircraft handling, estimated landing and take-off times for aircraft, delays in aircraft landings and take-offs, take-off and arrival separations, etc.
- Step 2.** Generating an initial population. Chromosomes (matrix structures) set the routes of aircraft movement on the apron, both take-off and landing routes.
- Step 3.** Setting the input parameters of the genetic algorithm i.e. number of iterations, population size, crossover and mutation parameters. The setting of the input parameters determines the correctness of the result generation.
- Step 4.** Each individual in the population is assessed according to its adaptation function. In the case under consideration, the evaluation function is the time of airport apron occupancy by aircraft, measured from landing to take-off (taxiing time).
- Step 5.** Using the roulette method, individuals with the best adaptation function are selected for the next generation (iteration of the algorithm).
- Step 6.** The process of the algorithm rapidly aiming at undesirable local minima blocked by the introduction of a scaling process.
- Step 7.** The purpose of the crossover process is to trigger genetic changes in a population of individuals to introduce new chromosomes into the population.
- Step 8.** The purpose of the mutation process is to trigger genetic changes in a population of individuals to introduce new chromosomes into the population.
- Step 9.** The repair algorithm is triggered in the case of an erroneous structure generated after the crossover and mutation process.
- Step 10.** Generating a final population about the interpretation of aircraft routing.

Steps 3-9 of the algorithm are repeated a specified number of iterations until a stop condition is obtained. The stop condition is a certain number of iterations. The matrix structure determines the routes of the aircraft movement on the apron. The matrix structure of the chromosome was randomly generated according to developed algorithms. The matrix structure has an interpretation of the decision variables developed in the mathematical model. The initial population consists of a certain number of matrix structures determined at the beginning of the algorithm.

The algorithm for selecting chromosomes for crossover takes into account the whole process of selecting chromosomes for crossover, in the case of chromosome oddity it randomly selects the chromosome to pair, randomly pairs the two chromosomes, randomly selects the cutting points of the chromosomes and activates the crossover algorithm adequate to the proposed matrix structure. The crossover algorithm is supported by an individual repair algorithm. The mutation algorithm draws the chromosome for the mutation process and swaps the values of randomly selected genes. The crossover and mutation algorithms occur with a certain probability defined as input data. The end result of the genetic algorithm is a generated population that determines a comprehensive set of aircraft movement routes on the apron. The parameters of the genetic algorithm i.e. crossover and mutation probabilities, number of iterations and population size were chosen experimentally. The process of verifying the genetic algorithm was carried out on the basis of comparison of the genetic algorithm solutions with those

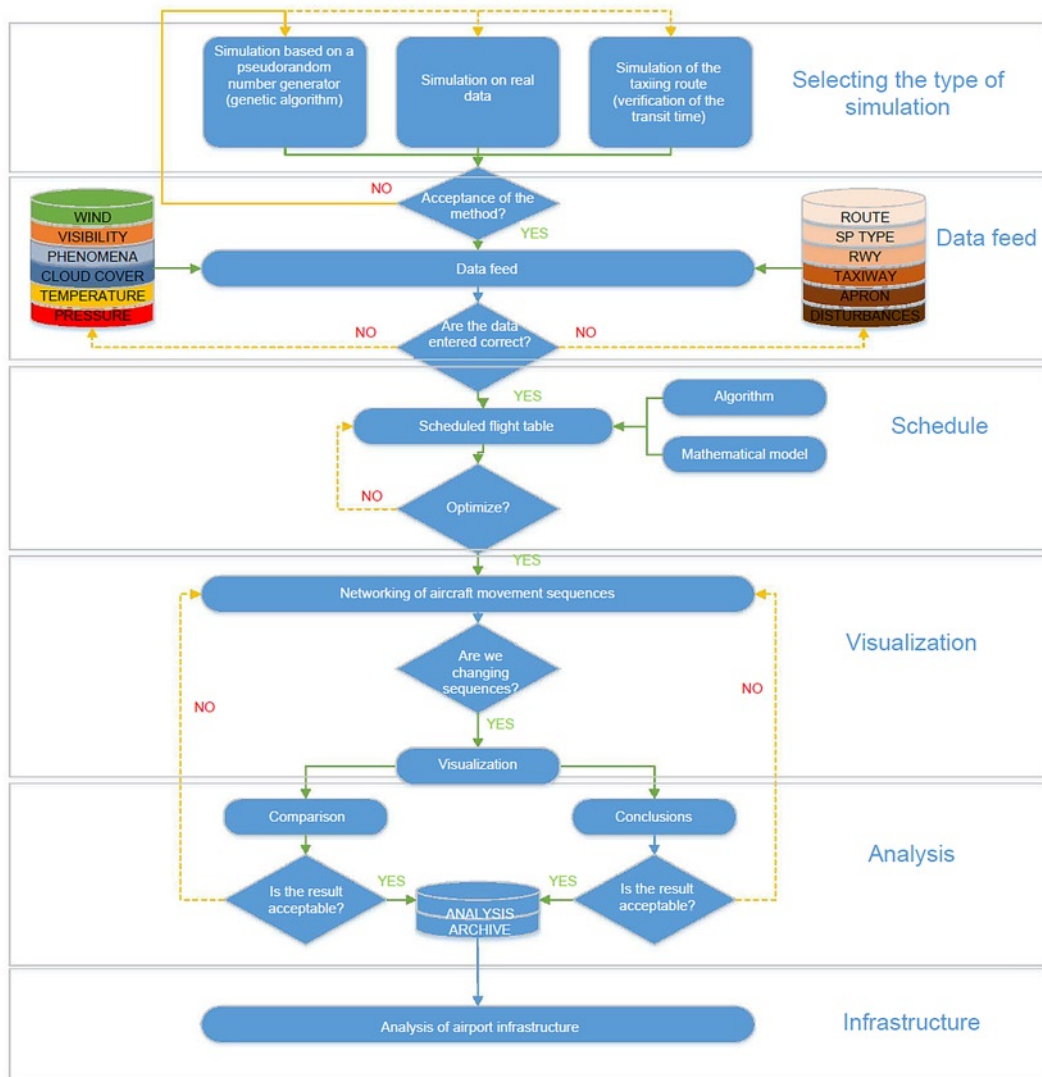


Fig. 6. Functional modules of the simulator

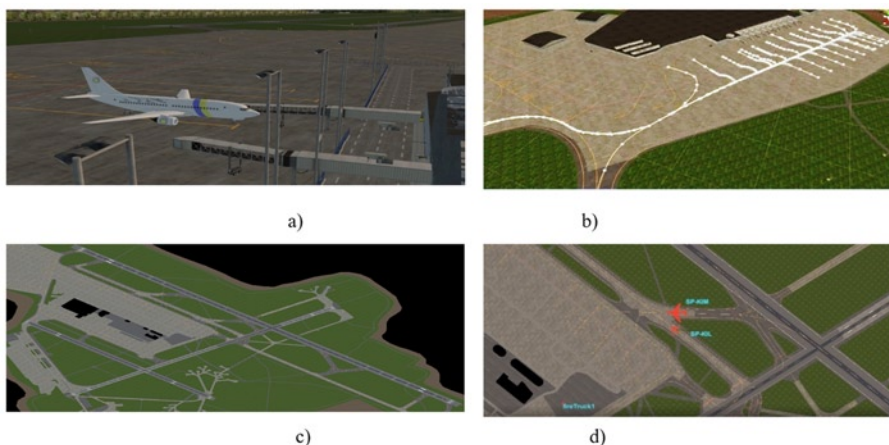


Fig. 7. Graphic representation of the simulation visualization module
Source: printout of the simulation using the simulation tool (own elaboration)

initial population (initial aircraft transit routes) for the genetic algorithm. This simulation is determined based on optimization processes so it is an effective tool for assessing the quality of airport processes.

The data feed module is used to enter various types of data such as service time of a given carrier and types of aircraft operated at a given airport. This data may also include the number of runways (RWY) or apron parking areas.

The scheduled flight table is an element that shows the arrival and departure times of aircraft from a given airport based on the data entered. This module is a kind of a schedule of the simulation set, thanks to which it is certain that given operations are planned and introduced correctly with simulation assumptions created on the basis of real data or random number generator.

The visualization module depicts the complete airport environment, including: depiction of aircraft (Fig. 7a), configuration of runways and taxiways (Fig. 7b), airport apron (Fig. 7c), natural terrain surrounding the airport, man-made objects or runway conditions.

The result analysis module verifies the correctness of the simulation based on the genetic algorithm. In case of an erroneous verification, the algorithm is calibrated by changing the initial settings of the algorithm parameters. Additionally, the performance analysis module evaluates the efficiency of a given airport in terms of existing airport infrastructure.

6. Practical example of using a simulation tool to evaluate the efficiency of traffic management processes on the apron

6.1. Simulation assumptions

To analyse and evaluate the organization of aircraft traffic on the apron, simulation studies were conducted on real data obtained from the operation of a real airport. The simulation assumed that different “S” and “D” taxiways would be used, and assumed that aircraft have different technical parameters (e.g., ground speed, acceleration time, braking time, taxiing time, and handling time).

Table 3. Aircraft taxiing times on the ways

| Measurement No. | Aircraft type | Taxiway marking | Taxiing time (min) | Measurement No. | Aircraft type | Taxiway marking | Taxiing time (min) |
|-----------------|---------------|-----------------|--------------------|-----------------|---------------|-----------------|--------------------|
| 1 | ATR72 | SOMZA32 | 2.06 | 37 | ATR72 | DAW76 | 1.56 |
| 2 | ATR72 | SOMZA32 | 1.59 | 38 | ATR72 | DAW76 | 1.35 |
| 3 | ATR72 | SOMZA32 | 1.35 | 39 | ATR72 | DAW76 | 1.29 |
| 4 | B737 | SA51 | 5 | 40 | ER145 | DA33 | 2.54 |
| 5 | B737 | SA51 | 4.58 | 41 | ER145 | DA33 | 2.35 |
| 6 | B737 | SA51 | 4.21 | 42 | ER145 | DA33 | 2.59 |
| 7 | MD87 | SOM24 | 2.55 | 43 | AVRO | DA36P | 3 |
| 8 | MD87 | SOM24 | 2 | 44 | AVRO | DA36P | 2.59 |
| 9 | MD87 | SOM24 | 2.22 | 45 | AVRO | DA36P | 3.19 |
| 10 | MD82 | SOM70 | 2.26 | 46 | B737 | DAZM12 | 3.38 |
| 11 | MD82 | SOM70 | 2.15 | 47 | B737 | DAZM12 | 3.29 |
| 12 | MD82 | SOM70 | 2.18 | 48 | B737 | DAZM12 | 3.41 |
| 13 | JS32 | SAW87 | 4.13 | 49 | B767 | DAZ10L | 2.28 |
| 14 | JS32 | SAW87 | 4.25 | 50 | B767 | DAZ10L | 2.25 |
| 15 | JS32 | SAW87 | 4.33 | 51 | B767 | DAZ10L | 3 |
| 16 | A320 | SOM11 | 8 | 52 | ER190 | DAZM32 | 2.3 |
| 17 | A320 | SOM11 | 8.36 | 53 | ER190 | DAZM32 | 2.1 |
| 18 | A320 | SOM11 | 9.05 | 54 | ER190 | DAZM32 | 2.45 |
| 19 | A321 | SOMZ10 | 2.38 | 55 | ATR72 | DAZM31 | 7 |
| 20 | A321 | SOMZ10 | 2.24 | 56 | ATR72 | DAZM31 | 6.54 |
| 21 | A321 | SOMZ10 | 2.17 | 57 | ATR72 | DAZM31 | 6 |
| 22 | CRJ | SOM35 | 1.15 | 58 | FOCKER | DAZM35 | 3.15 |
| 23 | CRJ | SOM35 | 1.21 | 59 | FOCKER | DAZM35 | 3.28 |
| 24 | CRJ | SOM35 | 1.36 | 60 | FOCKER | DAZM35 | 3.18 |
| 25 | ER180 | SOM14P | 2.21 | 61 | CRJ | DA34 | 3.1 |
| 26 | ER180 | SOM14P | 2.47 | 62 | CRJ | DA34 | 3.12 |
| 27 | ER180 | SOM14P | 2.14 | 63 | CRJ | DA34 | 3.06 |
| 28 | A319 | SOM13L | 2.45 | 64 | ER170 | DAZM21 | 4.3 |
| 29 | A319 | SOM13L | 2.15 | 65 | ER170 | DAZM21 | 4.28 |
| 30 | A319 | SOM13L | 3 | 66 | ER170 | DAZM21 | 4.56 |
| 31 | A319 | SOM19 | 3.56 | 67 | B737 | DAE48 | 6.29 |
| 32 | A319 | SOM19 | 3.48 | 68 | B737 | DAE48 | 6.45 |
| 33 | A319 | SOM19 | 3.23 | 69 | B737 | DAE48 | 6.18 |
| 34 | B737 | SOMZU5 | 5.42 | 70 | ER145 | DA33 | 2.54 |
| 35 | B737 | SOMZU5 | 5.3 | 71 | ER145 | DA33 | 2.59 |
| 36 | B737 | SOMZU5 | 6.01 | 72 | ER145 | DA33 | 2.38 |

Source: own study

Table 4. Taxiing times for aircraft on the “S” fast exit roads

| Type of Aircraft | Taxiway | Taxiing time [min] | Taxiing time according to simulation [min] | Difference |
|------------------|---------|--------------------|--|------------|
| ATR72 | SOMZA32 | 2.06 | 1.35 | -34% |
| ATR72 | SOMZA32 | 1.59 | 1.35 | -15% |
| ATR72 | SOMZA32 | 1.35 | 1.35 | 0% |
| B737 | SA51 | 5 | 4.15 | -17% |
| B737 | SA51 | 4.58 | 4.15 | -9% |
| B737 | SA51 | 4.21 | 4.15 | -1% |
| MD87 | SOM24 | 2.55 | 2.1 | -18% |
| MD87 | SOM24 | 2 | 2.1 | 5% |
| MD87 | SOM24 | 2.22 | 2.1 | -5% |
| MD82 | SOM70 | 2.26 | 2.1 | -7% |
| MD82 | SOM70 | 2.15 | 2.1 | -2% |
| MD82 | SOM70 | 2.18 | 2.1 | -4% |
| JS32 | SAW87 | 4.13 | 4.2 | 2% |
| JS32 | SAW87 | 4.25 | 4.2 | -1% |
| JS32 | SAW87 | 4.33 | 4.2 | -3% |
| A320 | SOM11 | 8 | 7.3 | -9% |
| A320 | SOM11 | 8.36 | 7.3 | -13% |
| A320 | SOM11 | 9.05 | 7.3 | -19% |
| A321 | SOMZ10 | 2.38 | 2.15 | -10% |
| A321 | SOMZ10 | 2.24 | 2.15 | -4% |
| A321 | SOMZ10 | 2.17 | 2.15 | -1% |
| CRJ | SOM35 | 1.15 | 1.1 | -4% |
| CRJ | SOM35 | 1.21 | 1.1 | -9% |
| CRJ | SOM35 | 1.36 | 1.1 | -19% |
| ER180 | SOM14P | 2.21 | 2.12 | -4% |
| ER180 | SOM14P | 2.47 | 2.12 | -14% |
| ER180 | SOM14P | 2.14 | 2.12 | -1% |
| A319 | SOM13L | 2.45 | 2.2 | -10% |
| A319 | SOM13L | 2.15 | 2.2 | 2% |
| A319 | SOM13L | 3 | 2.2 | -27% |
| A319 | SOM19 | 3.56 | 3.12 | -12% |
| A319 | SOM19 | 3.48 | 3.12 | -10% |
| A319 | SOM19 | 3.23 | 3.12 | -3% |
| B737 | SOMZU5 | 5.42 | 5.29 | -2% |
| B737 | SOMZU5 | 5.3 | 5.29 | 0% |
| B737 | SOMZU5 | 6.01 | 5.29 | -12% |

Source: own report based on data from the airport

The purpose of conducting the simulation is to compare the actual aircraft taxiing times with the taxiing times in the simulation environment using the optimization algorithm proposed in this paper. Tab. 3 shows actual aircraft taxiing times at the airport selected for the study.

6.2. Comparison of results

Aircraft movement studies using the simulation tool provided a percentage representation of the differences between actual taxiing times and times generated by the simulation process. The time gains when applying the simulation method in several cases reach or exceed 20%, which proves the high efficiency of the tool used and the cor-

rect verification of the optimization algorithm. The results of the percentage summary are presented sequentially in Tab. 4 for the taxiway starting from the “S” fast exit road (sierra) and Tab. 5 for the taxiway starting from the “D” fast exit road (delta).

7. Conclusions

The movement of aircraft on the apron must be based on well-considered decisions, taking into account many aspects of scheduling and finding the best route, in order to reduce overall travel times and to match the take-off and landing windows of individual aircraft to minimise the risk of potential collisions.

Table 5. Taxiing times for aircraft on "D" fast exit roads

| Type Aircraft | Taxiway | Taxiing time [min] | Taxiing time according to simulation [min] | Difference |
|---------------|---------|--------------------|--|------------|
| ATR72 | DAW76 | 1.56 | 1.2 | -23% |
| ATR72 | DAW76 | 1.35 | 1.2 | -11% |
| ATR72 | DAW76 | 1.29 | 1.2 | -7% |
| ER145 | DA33 | 2.54 | 2.1 | -17% |
| ER145 | DA33 | 2.35 | 2.1 | -11% |
| ER145 | DA33 | 2.59 | 2.1 | -19% |
| AVRO | DA36P | 3 | 3 | 0% |
| AVRO | DA36P | 2.59 | 3 | 16% |
| AVRO | DA36P | 3.19 | 3 | -6% |
| B737 | DAZM12 | 3.38 | 3.15 | -7% |
| B737 | DAZM12 | 3.29 | 3.15 | -4% |
| B737 | DAZM12 | 3.41 | 3.15 | -8% |
| B767 | DAZ10L | 2.28 | 2.12 | -7% |
| B767 | DAZ10L | 2.25 | 2.12 | -6% |
| B767 | DAZ10L | 3 | 2.12 | -29% |
| ER190 | DAZM32 | 2.3 | 2.11 | -8% |
| ER190 | DAZM32 | 2.1 | 2.11 | 0% |
| ER190 | DAZM32 | 2.45 | 2.11 | -14% |
| ATR72 | DAZM31 | 7 | - | -100% |
| ATR72 | DAZM31 | 6.54 | - | -100% |
| ATR72 | DAZM31 | 6 | - | -100% |
| FOCKER | DAZM35 | 3.15 | 3.18 | 1% |
| FOCKER | DAZM35 | 3.28 | 3.18 | -3% |
| FOCKER | DAZM35 | 3.18 | 3.18 | 0% |
| CRJ | DA34 | 3.1 | 2.59 | -16% |
| CRJ | DA34 | 3.12 | 2.59 | -17% |
| CRJ | DA34 | 3.06 | 2.59 | -15% |
| ER170 | DAZM21 | 4.3 | 4.11 | -4% |
| ER170 | DAZM21 | 4.28 | 4.11 | -4% |
| ER170 | DAZM21 | 4.56 | 4.11 | -10% |
| B737 | DAE48 | 6.29 | 6 | -5% |
| B737 | DAE48 | 6.45 | 6 | -7% |
| B737 | DAE48 | 6.18 | 6 | -3% |
| ER145 | DA33 | 2.54 | 2.3 | -9% |
| ER145 | DA33 | 2.59 | 2.3 | -11% |
| ER145 | DA33 | 2.38 | 2.3 | -3% |

Source: own report based on data from the airport

The research presented in this paper has confirmed the efficiency of a simulation tool based on a genetic algorithm used to evaluate airport processes.

The proposed simulation tool allows the analysis and evaluation of airport processes in the context of, among others: increasing airport capacity, planning the positioning of aircraft on the apron, extending taxiways, selecting the number of runways, optimizing aircraft taxiways on the apron, determining the order of take-offs and landings,

increasing the efficiency and effectiveness of airport processes in the context of safety of airport operations.

The application in the form of a simulation tool allows for the verification of the operation of a given airport in a given time interval.

The developed proprietary tool additionally enables the analysis and evaluation of operations related to take-off, landing, taxiing and handling of aircraft in real traffic conditions.

References

1. Adacher L, Flamini M, Romano E. Airport Ground Movement Problem: Minimization of Delay and Pollution Emission. IEEE Transactions on Intelligent Transportation Systems 2018; 19(12): 3830-3839, <https://doi.org/10.1109/TITS.2017.2788798>.

2. Anagnostakis I, Clarke JP, Böhme D, Völckers U. Runway operations planning and control: Sequencing and scheduling. *Journal of Aircraft* 2001; 38(6): 988-996, <https://doi.org/10.2514/2.2882>.
3. Anderson R, Milutinović D. An approach to optimization of airport taxiway scheduling and traversal under uncertainty. *Proceedings of the Institution of Mechanical Engineers Part G Journal of Aerospace Engineering* 2013; 227(2): 273-284, <https://doi.org/10.1177/0954410011433238>.
4. Atkin JAD, Burke EK, Greenwood JS, Reeson D. Hybrid metaheuristics to aid runway scheduling at London Heathrow airport. *Transportation Science* 2007; 41(1): 90-106, <https://doi.org/10.1287/trsc.1060.0163>.
5. Bianco L, Dell'Olmo P, Giordani S. Scheduling models for air traffic control in terminal areas. *Journal of Scheduling* 2006; 9: 223-253, <https://doi.org/10.1007/s10951-006-6779-7>.
6. Borucka A, Wiśniewski P, Mazurkiewicz D, Świderski A. Laboratory measurements of vehicle exhaust emissions in conditions reproducing real traffic. *Measurement* 2021; 174: 1-12, <https://doi.org/10.1016/j.measurement.2021.108998>.
7. Clare G, Richards AG. Optimization of taxiway routing and runway scheduling. *IEEE Transactions on Intelligent Transportation Systems* 2011; 12(4): 1000-1013, <https://doi.org/10.1109/TITS.2011.2131650>.
8. Daniewski K, Kosicka E, Mazurkiewicz D. Analysis of the correctness of determination of the effectiveness of maintenance service actions. *Management and Production Engineering Review* 2018; 9(2): 20-25, <https://doi.org/10.24425/119522>.
9. Dorndorf U, Drexl A, Nikulin Y, Pesch E. Flight gate scheduling: state-of-the-art and recent developments. *Omega* 2007; 35(3): 326-334, <https://doi.org/10.1016/j.omega.2005.07.001>.
10. Egbelu PJ, Tanchoko JMA. Characterization of automatic guided vehicle dispatching rules. *International Journal of Production Research* 1984; 22(3): 359-374, <https://doi.org/10.1080/00207548408942459>.
11. Eun Y, Hwang I, Bang H. Optimal arrival flight sequencing and scheduling using discrete airborne delays. *IEEE Transactions on Intelligent Transportation Systems* 2010; 11(2): 359-373, <https://doi.org/10.1109/TITS.2010.2044791>.
12. Garcia J, Berlanga A, Molina JM, Casar JR. Optimization of airport ground operations integrating genetic and dynamic flow management algorithms. *AI Communications* 2005; 18(2): 143-164.
13. Gołda P, Izdebski M, Szczepański E. The application of ant algorithm in the assignment problem of aircrafts to stops points on the apron. *Journal of KONES Powertrain and Transport* 2018; 25(1): 111-118, <https://doi.org/10.5604/01.3001.0012.7845>.
14. Gołda P, Kowalski M, Wasser C, Dygnatowski P, Szporka A. Elements of the model positioning of aircraft on the apron. *Archives of Transport* 2019; 51(3): 101-108, <http://dx.doi.org/10.5604/01.3001.0013.6166>.
15. Gołda P, Manerowski J. Support of aircraft taxiing operations on the apron. *Journal of KONES Powertrain and Transport* 2014; 21(4): 127-135, <https://doi.org/10.5604/12314005.1130457>.
16. Gołda P. Selected decision problems in the implementation of airport operations. *Scientific Journal of Silesian University of Technology. Series Transport* 2018; 101: 79-88, <https://doi.org/10.20858/sjsutst.2018.101.8>.
17. Herrero JG, Berlanga A, Molina JM, Casar JR. Methods for operations planning in airport decision support systems. *Applied Intelligence* 2005; 22(3): 183-206, <https://doi.org/10.1007/s10791-005-6618-z>.
18. Hockaday SLM, Kanafani AK. Developments in airport capacity analysis. *Transportation Research* 1974; 8: 171-180, [https://doi.org/10.1016/0041-1647\(74\)90004-5](https://doi.org/10.1016/0041-1647(74)90004-5).
19. Hoshino S, Seki H, Naka Y. Pipeless batch plant with operating robots for a multiproduct production system. *Distributed Autonomous Robotic Systems* 1987; 4: 33-51, https://doi.org/10.1007/978-3-642-00644-9_44.
20. Izdebski M, Jacyna M. The organization of the municipal waste collection: The decision model. *Annual Set The Environment Protection* 2018; 20: 919-933.
21. Izdebski M, Jacyna-Gołda I, Gołębiowski P, Plandor J. The Optimization Tool Supporting Supply Chain Management in the Multi-Criteria Approach. *Archives of Civil Engineering* 2020; 66(3): 505-524, <https://doi.org/10.24425/ace.2020.134410>.
22. Izdebski M, Jacyna-Gołda I, Jakowlewa I. Planning International Transport Using the Heuristic Algorithm. *Advances in Intelligent Systems and Computing* 2019; 844: 229-241, https://doi.org/10.1007/978-3-319-99477-2_21.
23. Izdebski M, Jacyna-Gołda I, Markowska K, Murawski J. Heuristic algorithms applied to the problems of servicing actors in supply chains. *Archives of Transport* 2017; 44(4): 25-34, <https://doi.org/10.5604/01.3001.0010.6159>.
24. Jacyna M, Izdebski M, Szczepański E, Gołda P. The Task Assignment of Vehicles for a Production Company. *Symmetry* 2018; 10(11): 551, <https://doi.org/10.3390/sym10110551>.
25. Jacyna M, Jachimowski R, Szczepański E, Izdebski M. Road vehicle sequencing problem in a railroad intermodal terminal-simulation research. *Bulletin of the Polish Academy of Sciences: Technical Sciences* 2020; 68(4): 1135-1148, <http://dx.doi.org/10.24425/bpasts.2020.134643>.
26. Jacyna M, Semenov I. Models of vehicle service system supply under information uncertainty. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2020; 22(4): 694-704, <http://dx.doi.org/10.17531/ein.2020.4.1>.
27. Jacyna-Gołda I, Izdebski M, Podwieszko A. Assessment of the efficiency of the assignment of vehicles to tasks in supply chains: A case-study of a municipal company. *Transport* 2017; 32(3): 243-251, <https://doi.org/10.3846/16484142.2016.1275040>.
28. Jiang Y, Xu X, Zhang H, Luo Y. Taxiing route scheduling between taxiway and runway in hub airport. *Mathematical Problems in Engineering* 2014; 2015(1): 1-14, <https://doi.org/10.1155/2015/925139>.
29. Kariya Y, Yahagi H, Takehisa M, Yoshihara S, Ogata T, Hara T, Ota J. Modeling and designing aircraft taxiing patterns for a large airport. *Advanced Robotics* 2013; 27(14): 1059-1072, <http://dx.doi.org/10.1080/01691864.2013.805469>.
30. Kowalski M, Izdebski M, Żak J, Gołda P, Manerowski J. Planning and management of aircraft maintenance using a genetic algorithm. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2021; 23(1): 143-153, <http://dx.doi.org/10.17531/ein.2021.1.15>.
31. Kuchar JK, Yang LC. A review of conflict detection and resolution modeling methods. *IEEE Transactions on Intelligent Transportation Systems* 2000; 4(1): 179-189, <https://doi.org/10.1109/6979.898217>.
32. Liu Y. Study on optimization for taxiway routing arrangement based on simulation. *Applied Mechanics and Materials* 2011; 97-98: 550-553, <https://doi.org/10.4028/www.scientific.net/AMM.97-98.550>.
33. Liu YH. A genetic local search algorithm with a threshold accepting mechanism for solving the runway dependent aircraft landing problem. *Optimization Letters* 2011; 5(2): 229-245, <https://doi.org/10.1007/s11590-010-0203-0>.
34. Mann GW, Hwang I. Four-dimensional aircraft taxiway conformance monitoring with constrained stochastic linear hybrid systems. *Journal of Guidance, Control, and Dynamics* 2012; 35(5): 1593-1604, <https://doi.org/10.2514/1.54438>.

35. Montoya J, Wood Z, Rathinam S. Runway Scheduling Using Generalized Dynamic Programming. American Institute of Aeronautics and Astronautics, AIAA Guidance, Navigation, and Control Conference Portland 2011, <https://doi.org/10.2514/6.2011-6380>.
36. Mori R. Aircraft ground-taxiing model for congested airport using cellular automata. IEEE Transactions on Intelligent Transportation Systems 2013; 14(1): 180-188, <https://doi.org/10.1109/TITS.2012.2208188>.
37. Nogueira KB, Aguiar PHC, Weigang L. Using ant algorithm to arrange taxiway sequencing in airport. International Journal of Computer Theory and Engineering 2014; 6(4): 357-361, <https://doi.org/10.7763/IJCTE.2014.V6.889>.
38. Nowakowski T. Reliability model of combined transportation system. [in:] Spitzer C, Schmocker U, Dang V N (ed.). Probabilistic Safety Assessment and Management. London: Springer 2004; 2012-2017, https://doi.org/10.1007/978-0-85729-410-4_323.
39. Pitfield DE, Jerrard EA. Monte Carlo comes Rome: a note on the estimation of unconstrained runway capacity at Rome fiumucino international airport. Journal of Air Transport Management 1999; 5: 185-192, [https://doi.org/10.1016/S0969-6997\(99\)00012-5](https://doi.org/10.1016/S0969-6997(99)00012-5).
40. Rathinam S, Montoya J, Jung Y. An optimization model for reducing aircraft taxi times at the Dallas Fort Worth International Airport. 26th International Congress of the Aeronautical Sciences 2008.
41. Ravizza S, Atkin JAD, Maathuis MH, Burke EK. A combined statistical approach and ground movement model for improving taxi time estimations at airports. Journal of the Operational Research Society 2013; 64(9): 1347-1360, <https://doi.org/10.1057/jors.2012.123>.
42. Rodríguez-Sanz Á, Fernández BR, Comendador FG, Valdés RA, García JMC, Bagamanova M. Operational Reliability of the Airport System: Monitoring and Forecasting. Transportation Research Procedia 2018; 33: 363-370, <https://doi.org/10.1016/j.trpro.2018.11.002>.
43. Roling PC, Visser HG. Optimal airport surface traffic planning using mixed-integer linear programming. International Journal of Aerospace Engineering 2008; 2008(1): 1-11, <https://doi.org/10.1155/2008/732828>.
44. Siddiquee W. A mathematical model for predicting the number of potential conflict situations at intersecting air routes. Transportation Science 1973; 7: 158-167, <https://doi.org/10.1287/trsc.7.2.158>.
45. Solveling G, Solak S, Clarke JP, Johnson E. Runaway operations optimization in the presence of uncertainties. Journal of Guidance Control, and Dynamics 2011; 34(5): 1373-1381, <https://doi.org/10.2514/1.52481>.
46. Szaciłło L, Szczepański E, Izdebski M, Jacyna M. Risk assesment for rail freight transport operations. Eksploatacja i Niezawodność – Maintenance and Reliability 2021; 23(3): 476-488, <http://doi.org/10.17531/ein.2021.3.8>.
47. Turskis Z, Antuchevičienė J, Keršulienė V, Gaidukas G. Hybrid Group MCDM Model to Select the Most Effective Alternative of the Second Runway of the Airport. Symmetry 2019; 11(6):792. <https://doi.org/10.3390/sym11060792>.
48. Wasiak M, Jacyna-Gołda I, Markowska K, Jachimowski R, Kłodawski M, Izdebski M. The use of a supply chain configuration model to assess the reliability of logistics processes. Eksploatacja i Niezawodność – Maintenance and Reliability 2019; 21 (3): 367–374, <http://dx.doi.org/10.17531/ein.2019.3.2>.
49. Zieja M, Smoliński H, Gołda P. Information systems as a tool for supporting the management of aircraft flight safety. Archives of Transport 2015; 36(4): 67-76, <http://dx.doi.org/10.5604/08669546.1185211>.

Article citation info:

Nguyen T-P, Lin Y-K. Investigation of the influence of transit time on a multistate transportation network in tourism. *Eksploracja i Niezawodność – Maintenance and Reliability* 2021; 23 (4): 670–677, <http://doi.org/10.17531/ein.2021.4.9>.

Investigation of the influence of transit time on a multistate transportation network in tourism

Thi-Phuong Nguyen^a, Yi-Kuei Lin^{b,c,d,e,*}

^a Department of Distribution Management, National Chin-Yi University of Technology, Taichung 411, Taiwan

^b Department of Industrial Engineering & Management, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan

^c Department of Business Administration, Asia University, Taichung 413, Taiwan

^d Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404, Taiwan

^e Department of Business Administration, Chaoyang University of Technology, Taichung 413, Taiwan

Indexed by:



Highlights


- Propose an algorithm to investigate how transit times affect the performance of an MTN.
- A performance indicator termed “reliability” is used as a basis for the investigation.
- The sufficient levels of transit time are determined to examine.
- Provide management suggestions about appropriate transit times to maintain reliability.

Abstract

Abstract

Reliability has been widely used as a potential indicator of the performance assessment for several real-life networks. Focus on a multistate transportation network in tourism (MTN), this study evaluates the reliability of the MTN as a basis for investigating the influence of transit time. Reliability is the probability to fulfill transportation demand under the given time threshold and budget limitation and evaluated at various levels of transit times. An algorithm, which employs the boundary points and recursive sum of disjoint products technique, is proposed to evaluate the MTN reliability. According to the obtained results, this paper analyzes the influence of transit times on MTN reliability. Particularly, this paper discusses and provides some suggestions about the appropriate transit time to maintain reliability. Decision-makers in the tourism industry also can predetermine the significant drops of reliability to improve the relevant transit times. Besides, the proposed investigation is indicated and proved through an illustrative example and a practical case.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

multistate transportation network, reliability, transit times, budget limitation, time threshold.

1. Introduction

A transportation network, which combines various modes of transport such as sea, air, road, and rail, becomes more popular and is applied in many systems [1, 8, 11, 14, 20, 23, 28, 34]. Toward environmental and economic sustainability, decision-makers in logistics management often consider trains, trucks, and barges to design their multimodal transportation networks [14, 34]. As a crucial part, a transportation network contributes to thriving travel agents who business the tourism industry [2, 3, 30]. Besides, maintaining service quality stable and reliable is vital from the management perspective in most service industries. Thus, a reliable transportation network can efficiently complete operational functions and smoothly provide customers high-quality tourism services. It raises a need to evaluate the performance of transportation networks in tourism. In recent decades, reliability, which is the ability to complete requested functions/ tasks under given constraints in a predetermined period, is an appropriate and widely used performance indicator [7]. In terms of connectivity

performance, reliability has been defined as the probability that the source can link with the sink [10, 31]. Concerning the terms “flow” and “capacity”, the ability to fulfill a required demand is considered as reliability [6, 9, 11, 25]. For instance, reliability has been studied as the probability that a logistic network can deliver a given volume of goods to a specific destination [11]. Considering on-time performance, Nguyen and Lin [25] measured reliability as the ability of an air transport network to successfully carry a given number of passengers to the final destinations within a specific time threshold. To address the reliability evaluation, various methods including cross-entropy [24], state enumeration [21], percolation theory [16, 19], and minimal cut-sets and path-sets [25, 26, 31, 32] have been proposed.

Furthermore, many studies consider time and budget, which are two of the key influencing factors in customer satisfaction and transportation choices [8, 15, 17, 18, 22, 27], when investigating the reliability of transportation networks. Survey the transport behavior during the COVID-19 pandemic, Das et al., indicated a significant

(*) Corresponding author.

E-mail addresses: T-P. Nguyen - phuongnt@ncut.edu.tw, Y-K. Lin - yklin@nctu.edu.tw

impact of some factors including monthly income and travel time on the transport switch of the participants [8]. Besides, the flights operated by low-cost carriers at Incheon international airport (Korea) reported a double increase during the year 2012 to 2015, which differs from a 10% growth of full-service carriers [12]. In the same report, the passenger market share of low-cost carriers increased from 5.7% in 2012 to 15.9% in 2015. This boom of low-cost carriers infers an attractiveness of price to customers [5, 13]. In developing countries, people with low income tend to have a higher frequency of using public transports [1]. This issue may be explained by the budget gap between the two groups. Regarding transportation time, the choice of customers may be affected by both the time to change routes if necessary and the total transportation time. For example, if the first choice for customers to arrive at the destination is taking a bus in one and a quarter-hour to the nearest airport thirty minutes before their one-hour flight. Another choice is riding a public bike in forty to the metro station in their area, walking around fifteen minutes to enter the metro line, then taking the two-hour metro to the destination. Clearly, the first choice takes ten minutes less time than the second one, but its transit time from bus to flight is a half-hour while that of the second choice is only fifteen minutes. Thus, customers may choose the second choice if they have enough time and do not want to check-in at the airport; otherwise, the first choice is a priority. Simply speaking, not only the total transportation time but also transit times (i.e., the required time to transfer between two routes in the tourism trip) are considered by customers. Since the transit time is the necessary time to take the next route, the transit time in the first choice is thirty minutes to walk from the bus station to the boarding gate of the flight, check-in, and customs check. In the second choice, the transit time infers the time that is around fifteen minutes for entering the metro line from the bicycle parking. Obviously, shortening the transit time can reduce the total transportation time and let customers enjoy journeys with multiple transport modes. However, not many studies appraise the effect on the transportation networks' reliability of the transit time. Therefore, this study targets to fill this academic gap.

The purpose of this study is to provide a new reliability-based approach to investigate the influence of transit time. Normally, a transportation network in tourism is a combination of air, sea, and road networks [23]. Like the studies in the literature [11, 20, 28], the transportation network in tourism is a typical flow-network that contains vertices (i.e., stations, airports, seaports) and directed edges (routes). Each route connecting a pair of stations is served by a particular vehicle and its capacity is the number of available seats. In general, the capacity is given and depends on the vehicle's size and design. However, some seats may be booked or reserved by others such as individual tourists and travel agents/ tours operators. That is, the capacity (available seats) of edges is multistate. Regarded as multistate capacity, a transportation network in tourism can be formulated as a multistate flow-network [32, 33, 35]. "A multistate transportation network (MTN)" is referred to as a transportation network in tourism herein. The MTN's reliability is calculated as the probability to meet transportation demand within a specific time threshold and budget limitation under different required transit times. To address the research problem, an algorithm, which employs the concept of boundary points, is proposed for evaluating reliability and analyzing the influence of transit times accordingly. Simultaneously, this study applies the recursive sum of disjoint product (RSDP) technique [4, 33] to compute reliability as the probability of the MTN's states demarcated by the boundary points. As a basis, the obtained reliability contributes to indicating the performance of the MTN and investigating the effect of transit time. In addition, appropriate transit times are provided towards a more reliable performance of the MTN.

2. MTN model

In this section, the constructed model of a multistate transportation network (MTN) is introduced first. An MTN is characterized by V – set of vertices (stations), E – set of directed edges (routes) e_j for $j = 1, 2, \dots, m$, G – set of transport costs g_j , and M – set of transport modes. Besides, the maximal capacity vector $C = (c_1, c_2, \dots, c_m)$ bounds all states (capacity vectors, Y) of the MTN. Thus, the maximal capacity c_j of each edge limits its current capacity y_j and flow $f(e_j)$. In short, the MTN is denoted as $N = (V, E, G, M, C)$. Each directed edge e_j in E is scheduled to move from its departure station d_j at td_j (departure time) to its arrival station a_j at ta_j (arrival time). Note that d_j and a_j belong to V . Besides, each edge uses a particular transport mode ($m_j \in M$) with a specific transport cost ($g_j \in G$). The remaining notations are listed below.

| | |
|---------------|---|
| s, t | source and sink, $\in V$ |
| $f(s, t)$ | flow from s to t |
| T | time threshold |
| B | budget limitation |
| A | transportation demand |
| w | transit time between the same modes |
| w^* | transit time between different modes |
| W | (w, w^*) : transit time vector |
| $W \leq W^i$ | i^{th} level of transit times |
| P_k | minimal path feasible under T and B (MTBP) |
| P | set of all MTBPs |
| U_k | (u, u^*) : maximal transit time vector that guarantees the validity of P_k |
| P^i | set of all feasible P_k if $W \leq W^i$ |
| $f(P_k)$ | flow through the minimal path $P_k \in P^i$ |
| F | $(f(P_k) P_k \in P^i)$: flow vector feasible under T and B with $W \leq W^i$ |
| F^i | set of all flow vectors meeting A under T and B with $W \leq W^i$ |
| Y | (y_1, y_2, \dots, y_m) : capacity vector |
| X^i | set of lower boundary point (LBP) candidates |
| L^i | set of lower boundary points of X^i |
| $R_{B,A,T}^i$ | reliability that N can meet transportation demand A under time threshold T and budget limitation B transit times $W \leq W^i$ |

Furthermore, the following are all assumptions made in this study.

- (i) The conservation law of flows is followed.
- (ii) The capacity of all routes is independent statistically.
- (iii) Transit times between the same and different transport modes are considered.
- (iv) All routes are on time.

3. Assess the MTN reliability

According to the research problem, it is not efficient to calculate the reliability of all transit times $W = (w, w^*)$, where w and w^* are the required time for transferring to the next route with the same transport mode and different transport mode, respectively. In this study, we consider different levels of transit times, $W \leq W^i$, and accordingly assess the MTN's reliability ($R_{B,A,T}^i$). Note that all transit times ($W \leq W^i$) have the same influence on reliability, and reliability is the probability to satisfy transportation demand A under the time threshold T and budget limitation B . In other words, $R_{B,A,T}^i$ is the probability that

the MTN can transport at least A passengers within T hours and the total transport cost does not exceed B with the required transit time $W \leq W^i$. Let Y be a capacity vector meeting A under T and B with $W \leq W^i$ and store it in Δ . Any capacity vectors satisfying the following condition belong to Δ .

Condition 1. Under time threshold T and budget limitation B with the required transit time $W \leq W^i$, there are at least A passengers transported. The MTN's reliability is the probability of all Y s in Δ , $R_{B,A,T}^i = \Pr\{\bigcup_{Y \in \Delta} \{Y | Y \leq C\}\}$. However, employing the concept of all upper and lower boundary points to calculate is more efficient than enumerating to determine the set Δ . In fact, both lower and upper boundary points are in Δ such that none in Δ is less than lower boundary points (LBP) and greater than upper boundary points (UBP). That means any Y in Δ is between at least one LBP and one UB. Note that the maximal capacity vector C is the only UB of the MTN and the LBPs (X) must meet not only Condition 1 but also Condition 2.

Condition 2. There exists no $Y \in \Delta$ that $Y \leq X$.

After determining all LBPs and storing them in a set L^i , the following formula is used to compute the reliability:

$$R_{B,A,T}^i = \Pr\{\bigcup_{X \in L^i} \{Y | X \leq Y \leq C\}\} \quad (1)$$

3.1. Minimal paths feasible under time threshold and transit times

To partly fulfill Condition 1, we first determine all minimal paths feasible under T and B (MTBPs). Namely, a MTBP is a sequence of edges that can link s to t within T hours and B with $W \leq W^i$ and do not visit any vertex twice. Assume that all MTBPs are stored in P^i . Each P_k in P^i must satisfy that:

- If e_j is the first edge and e_h is the last edge of P_k then
 - The edge e_j departs from the source ($d_j = s$).
 - The edge e_h arrives at the sink ($a_h = t$).
 - The transportation time on P_k does not exceed the time threshold ($ta_h - td_j \leq T$).
- Only edge e_j arrives at the departure station of e_h ($a_j = d_h$) with $W = (w, w^*)$ satisfying the following can connect to e_h .
 - If the transport mode of two edges is the same ($m_j = m_h$) then $td_h \geq ta_j + w$.
 - Otherwise, $td_h \geq ta_j + w^*$.
- The total transport cost on P_k does not exceed the limitation budget ($\sum_{e_j \in P_k} g_j \leq B$).

Let $f(P_k)$ be a flow through the minimal path P_k . Based on the conservation law and the MTN's capacity, the following constraints must be satisfied:

$$\begin{cases} f(e_j) = \sum_{e_j \in P_k} f(P_k) \\ f(e_j) \leq c_j \end{cases}, \text{ for } j = 1, 2, \dots, m. \quad (2)$$

Consequently, a flow vector $F = (f(P_k) | P_k \in P^i)$ meeting the time threshold T and budget limitation B with transit times $W \leq W^i$ is feasible under capacity Y if:

$$\sum_{e_j \in P_k} f(P_k) \leq y_j \leq c_j, \text{ for } j = 1, 2, \dots, m. \quad (3)$$

Like constraint (2), $f(s, t) = \sum_{P_k \in P^i} f(P_k)$. The remaining of Condition 1 becomes:

$$f(s, t) \geq A \text{ where } f(s, t) = \sum_{P_k \in P^i} f(P_k). \quad (4)$$

Hence, any capacity vector (X) is said to belong to Δ if its feasible flows (F) meet constraint (5).

$$\sum_{P_k \in P^i} f(P_k) \geq A, \text{ for } j = 1, 2, \dots, m. \quad (5)$$

3.2. Lower boundary points and reliability evaluation

From all capacity vectors (Y) above, Condition 2 is tested to obtain lower boundary points (X). Let F^i be a set of all flow vectors fulfilling the following constraint:

$$\sum_{P_k \in P^i} f(P_k) = A \quad (6)$$

Any capacity vector $X = (x_1, x_2, \dots, x_m)$ satisfying constraint (6) and the following constraint belongs to Δ . They are less than or equal to other capacity vectors $Y \in \Delta$ that $\sum_{e_j \in P_k} f(P_k) \leq y_j \leq c_j$ for at least one j or $\sum_{P_k \in P^i} f(P_k) > A$:

$$\sum_{e_j \in P_k} f(P_k) = x_j, \text{ for } j = 1, 2, \dots, \quad (7)$$

However, it is not sufficient for them to meet condition 2 because they may less than or equal to others. Hence, they are called lower boundary point candidates herein. Remark 1 indicates the features of an LBP candidate.

Remark 1. X is an LBP candidate if at least one F that satisfies constraints (6) and (7).

Let X^i store all LBP candidates. To gain exact LBPs in L^i , compare and remove the duplicates and the components that are greater than others from X^i . By applying the RSDP method [4, 35], the MTN's reliability can be easily derived through the formula (8):

$$R_{B,A,T}^i = \Pr\{\bigcup_{X \in L^i} \{Y | X \leq Y \leq C\}\} \quad (8)$$

3.3. Main algorithm to investigate the influence of transit time on the MTN reliability

The provided process describes how to evaluate the MTN reliability. However, to examine the effect of different transit times on reliability, evaluating reliabilities under all possible transit times is not efficient enough. This study employs the reliability evaluation and proposes an assessment algorithm under the budget limitation and time threshold. Firstly, suppose that the transit times are not required, $W = (0, 0)$, we determine all minimal paths (P_k) feasible under B and T then record them in a set P^0 . Simultaneously, obtain the maximal transit time $U_k = (u, u^*)$ – the validity condition of each P_k . Namely, each MTBP (P_k) is valid if $W \leq U_k$; otherwise, it is broken. A search procedure shown in Fig. 1 is developed to determine all MTBPs in P^i and their corresponding maximal transit time vectors.

Without considering the impact of transit times, the set P^0 contains all possible minimal paths of the MTN. And some of MTBPs in P^0 may be broken at a specific $W = (w, w^*)$ that $w > u$ and/ or $w^* > u^*$.

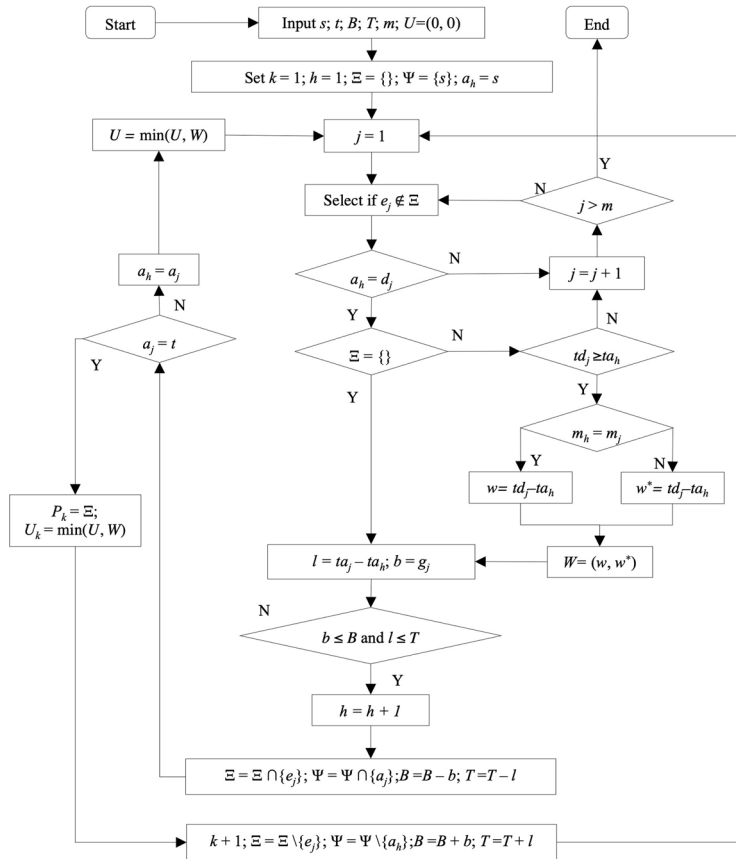


Fig. 1. Procedure to determine all MTBPs (P_k) and their maximal transit time vectors (U_k)

Thus, the set \mathbf{P}^i of all MTBPs in the case of existing transits times, $W \leq W^i$, is the sub-union of \mathbf{P}^0 . The MTN's reliability is impacted if and only if the $W \leq W^i$ can make at least one MTBP in \mathbf{P}^0 invalid (i.e., $\mathbf{P}^i \subset \mathbf{P}^0$). This research combines the values u and u^* of the same or different maximal transit times U_k to create W^i . Considering levels of transit time $W \leq W^i$ is sufficient for the study's analysis. The following algorithm is used to access the effect of transit time on reliability under the time threshold and budget limitation.

Main algorithm – Reliability assessment subject to the impact of transit times

Input: $N = (\mathbf{V}, \mathbf{E}, \mathbf{G}, \mathbf{M}, \mathbf{C}), T, B$, and A

Step 1: Apply the search procedure shown in Fig. 1 to generate \mathbf{P}^0 – set of all feasible minimal paths P_k under time threshold T and budget limitation B without required transit times. At the same time, obtain the corresponding maximal transit times U_k .

Step 2: From all maximal transit times U_k , create all possible W^i . Some $W^i = U_k$ and other $W^i = (w, w^*)$ where $w = u$ and $w^* = u^*$ of two different U_k .

Step 3: Conduct the following steps for each level of transit times $W \leq W^i$:

Step 3.1: Accept from \mathbf{P}^0 all minimal paths P_k such that $W^i \leq U_k$ and store them as \mathbf{P}^i .

Step 3.2: Determine all flow vectors $F = (f(P_k)) P_k \in \mathbf{P}^i$ that satisfy the following constraints to store as \mathbf{F}^i :

$$\sum_{P_k \in \mathbf{P}^i} f(P_k) = A, \text{ for } j = 1, 2, \dots, m. \quad (9)$$

Step 3.3: Through the following equation, convert from each F in \mathbf{F}^i to gain LBP candidates X and store in \mathbf{X}^i :

$$x_j = \sum_{e_j \in P_k} f(P_k), \text{ for } j = 1, 2, \dots, m. \quad (11)$$

Step 3.4: Compare all candidates in \mathbf{X}^i to remove the duplicates and the components that are greater than others. Exact LBPs are obtained as a set \mathbf{L}^i .

Step 3.5: By utilizing the RSDP method, compute the MTN's reliability as equations (12):

$$R_{B,A,T}^i = \Pr\left\{ \bigcup_{X \in \mathbf{L}^i} \{Y | X \leq Y \leq C\} \right\}. \quad (12)$$

Step 4: List all reliabilities under different levels of transit time in order of W^i .

4. Numerical example

This section introduces an MTN example that consists of four stations, eight routes, and three transport modes, shown in Fig. 2. Then, we demonstrate how to analyze the impacts of transit times on the MTN's reliability. The source is the first station, and the sink is the last station. The relevant data of eight routes in the MTN is shown in Table 1. After applying the main algorithm, the MTN's reliabilities to meet transportation demand $A = 80$ passengers under budget limitation $B = 200$ USD and time threshold $T = 8$ hours with different transit times are evaluated as follows:

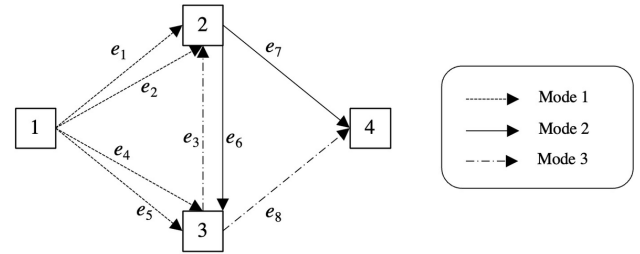


Fig. 2. An MTN example

Table 1. The relevant data about all routes in the MTN

| Route | Departure - Arrival time | Departure - Arrival station | Transport mode | Transport Cost (USD) |
|---------|--------------------------|-----------------------------|----------------|----------------------|
| (e_i) | $(d_j - a_j)$ | $(td_j - ta_j)$ | (g_i) | (g_i) |
| 1 | 8:00 – 9:00 | 1 – 2 | 1 | 45 |
| 2 | 7:45 – 11:40 | 1 – 2 | 1 | 25 |
| 3 | 10:00 – 11:30 | 3 – 2 | 3 | 50 |
| 4 | 8:15 – 9:30 | 1 – 3 | 1 | 50 |
| 5 | 8:15 – 12:05 | 1 – 3 | 1 | 25 |
| 6 | 9:45 – 11:15 | 2 – 3 | 2 | 45 |
| 7 | 12:00 – 15:30 | 2 – 4 | 2 | 100 |
| 8 | 12:30 – 16:00 | 3 – 4 | 3 | 110 |

Table 2. The capacity probability of all routes in the MTN

| Route (e_i) | Probability Pr (y_j passengers) | | | | | |
|-----------------|------------------------------------|---------|---------|---------|--------|-------|
| | 41 – 50 | 31 – 40 | 21 – 30 | 11 – 20 | 1 – 10 | 0 |
| 1 | 0.81 | 0.05 | 0.1 | 0.02 | 0.01 | 0.01 |
| 2 | 0.80 | 0.1 | 0.05 | 0.02 | 0.01 | 0.02 |
| 3 | | 0.93 | 0.05 | 0.01 | 0.005 | 0.005 |
| 4 | 0.81 | 0.05 | 0.06 | 0.05 | 0.02 | 0.01 |
| 5 | 0.80 | 0.08 | 0.05 | 0.05 | 0.01 | 0.01 |
| 6 | | 0.93 | 0.04 | 0.005 | 0.015 | 0.01 |
| 7 | 0.86 | 0.01 | 0.1 | 0.01 | 0.01 | 0.01 |
| 8 | 0.90 | 0.05 | 0.02 | 0.01 | 0.01 | 0.01 |

Input: $N = (V, E, G, M, C)$, $T = 8$ hours, $B = 200$ USD, and $A = 80$ passengers.

Step 1: Apply the search procedure shown in Fig. 1 to generate P^0 . In total, it contains six MTBPs in and the corresponding maximal transit times U_v are presented as below. Note that the unit of U_k is minute and the symbol “_” means that all transit times are accepted.

| | |
|---------------------------|-------------------|
| $P_1 = \{e_1, e_6, e_8\}$ | $U_1 = (45, _)$ |
| $P_2 = \{e_1, e_7\}$ | $U_2 = (180, _)$ |
| $P_3 = \{e_2, e_7\}$ | $U_3 = (_, 20)$ |
| $P_4 = \{e_4, e_3, e_7\}$ | $U_4 = (30, _)$ |
| $P_5 = \{e_4, e_8\}$ | $U_5 = (180, _)$ |
| $P_6 = \{e_5, e_8\}$ | $U_6 = (_, 25)$ |

Step 2: From all minimal transit times U_v , create and gain eight possibilities of W that are $W^1 = (30, 20)$, $W^2 = (30, 25)$, $W^3 = (45, 20)$, $W^4 = (45, 25)$, $W^5 = (180, 20)$, and $W^6 = (180, 25)$.

Step 3: Conduct the following steps for all levels of transit times from $W \leq W^1$ to $W \leq W^6$. For example, with the required transit time $W \leq (30, 25)$, the reliability is computed as follows.

Step 3.1: From P^0 , five minimal paths are accepted to get $P^2 = \{P_1, P_2, P_4, P_5, P_6\}$ because their $U_k \geq W^2$.

Step 3.2: Obtain 154 flow vectors $F = (f(P_1), f(P_2), f(P_4), f(P_5), f(P_6))$ that satisfy the following constraints and store them as F^2 :

$$\sum_{P_k \in P^2} f(P_k) = 80 \quad (13)$$

$$\begin{aligned} f(P_1) + f(P_2) &\leq c_1; \\ f(P_4) &\leq c_3; \\ f(P_4) + f(P_5) &\leq c_4; \\ f(P_6) &\leq c_5; \\ f(P_1) &\leq c_6; \\ f(P_2) + f(P_4) &\leq c_7; \\ f(P_1) + f(P_5) + f(P_6) &\leq c_8. \end{aligned} \quad (14)$$

Since $c_1 = c_4 = c_5 = c_7 = c_8 = 50$ and $c_3 = c_6 = 40$, constraint (14) can be shortened as follows:

$$\begin{aligned} f(P_1) + f(P_2) &\leq 50; \\ f(P_4) + f(P_5) &\leq 50; \\ f(P_1) &\leq 40; \\ f(P_2) + f(P_4) &\leq 50; \\ f(P_1) + f(P_5) + f(P_6) &\leq 50. \end{aligned} \quad (15)$$

Step 3.3: Convert from each F in F^2 through the following equation to gain candidates X and store in X^2 :

$$\begin{aligned} x_1 &= f(P_1) + f(P_2); \\ x_3 &= f(P_4); \\ x_4 &= f(P_4) + f(P_5); \\ x_5 &= f(P_6); \\ x_6 &= f(P_1); \\ x_7 &= f(P_2) + f(P_4); \\ x_8 &= f(P_1) + f(P_5) + f(P_6). \end{aligned} \quad (16)$$

Step 3.4: Compare 154 candidates stored in X^2 and remove the duplicates and the components that are greater than others. There are 67 exact LBPs obtained and recorded in a set L^2 .

Step 3.5: Through the RSDP method, compute the MTN's reliability as equations (17):

$$R_{200,80,8}^2 = \Pr\left\{\bigcup_{X \in L^2} \{Y \mid X \leq Y \leq C\}\right\} = 0.927964. \quad (17)$$

Step 4: All reliabilities under different levels of transit time from $W \leq W^1$ to $W \leq W^6$ are listed in the following figure.

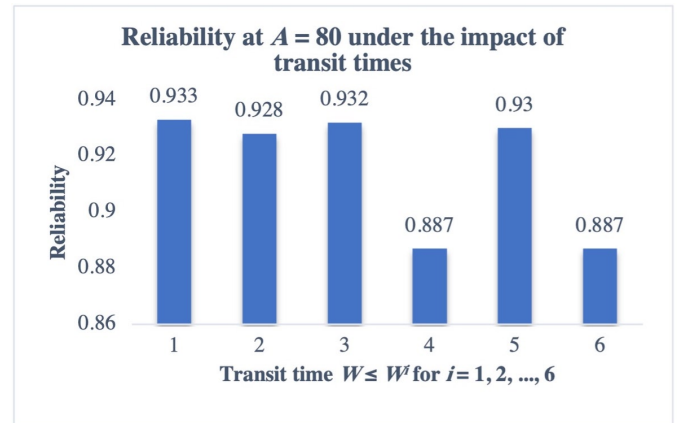


Fig. 3. The MTN's reliabilities at $A = 80$ passengers under the impact of transit times

As the results shown in Fig. 3, the MTN's reliability varies from 0.886554 to 0.933458. In which, it reaches a peak with transit times $W \leq \{W^1 \mid W^1 = (30, 20)\}$ and drops a bottom with $W \leq \{W^6 \mid W^6 = (180, 25)\}$. The reliability is higher than 0.8 at all cases of transit times and higher than 0.9 with four of six levels of transit times, which means that the ability to transport 80 passengers within 8 hours and 200 USD of the MTN is quite high. That means this MTN is quite reliable under the given time and budget limitations. When increasing the required transit time between the same modes, the reliability changes slightly with the required transit time between different transport modes $w^* \leq 20$; but it changes significantly with $w^* \leq 25$. At the same time, the

Table 3. The relevant data of the practical MTN

| Route | Dep. | Arr. | Dep. | Arr. | Mode | Cost |
|---------|----------|----------|---------|---------|---------|---------|
| | time | time | station | station | | (USD) |
| (e_i) | (td_i) | (ta_i) | (d_i) | (a_i) | (m_i) | (g_i) |
| 1 | 10:00 | 11:15 | CHU | TPE | HSR | 10 |
| 2 | 7:15 | 12:00 | CHU | TPE | Rail | 5 |
| 3 | 11:55 | 14:15 | TPE | HAN | Air | 95 |
| 4 | 19:45 | 22:00 | TPE | HAN | Air | 100 |
| 5 | 16:00 | 18:05 | HAN | HPH | Bus | 10 |
| 6 | 17:30 | 19:35 | HAN | HPH | Bus | 10 |
| 7 | 19:00 | 21:05 | HAN | HPH | Bus | 10 |
| 8 | 20:30 | 22:35 | HAN | HPH | Bus | 10 |
| 9 | 22:00 | 0:05 | HAN | HPH | Bus | 10 |
| 10 | 23:30 | 1:35 | HAN | HPH | Bus | 10 |
| 11 | 14:30 | 17:15 | TPE | DAD | Air | 115 |
| 12 | 17:45 | 18:40 | DAD | HPH | Air | 45 |
| 13 | 11:45 | 12:40 | DAD | HPH | Air | 45 |
| 14 | 9:10 | 12:00 | TPE | SGN | Air | 110 |
| 15 | 12:10 | 15:00 | TPE | SGN | Air | 110 |
| 16 | 15:10 | 18:00 | TPE | SGN | Air | 110 |
| 17 | 13:20 | 14:25 | SGN | HPH | Air | 55 |
| 18 | 15:20 | 15:30 | SGN | HPH | Air | 55 |
| 19 | 18:20 | 19:25 | SGN | HPH | Air | 55 |
| 20 | 9:05 | 9:20 | CHU | RMQ | Rail | 2 |
| 21 | 11:05 | 10:20 | CHU | RMQ | Rail | 2 |
| 22 | 12:05 | 11:20 | CHU | RMQ | Rail | 2 |
| 23 | 13:05 | 12:20 | CHU | RMQ | Rail | 2 |
| 24 | 14:05 | 13:20 | CHU | RMQ | Rail | 2 |
| 25 | 15:05 | 14:20 | CHU | RMQ | Rail | 2 |
| 26 | 16:05 | 15:20 | CHU | RMQ | Rail | 2 |
| 27 | 17:05 | 16:20 | CHU | RMQ | Rail | 2 |
| 28 | 15:00 | 18:15 | RMQ | HAN | Air | 115 |
| 29 | 13:15 | 16:45 | RMQ | DAD | Air | 120 |
| 30 | 10:00 | 11:15 | CHU | KHH | HSR | 10 |
| 31 | 7:15 | 12:00 | CHU | KHH | Rail | 5 |
| 32 | 9:10 | 12:00 | KHH | SGN | Air | 110 |
| 33 | 12:10 | 15:00 | KHH | SGN | Air | 110 |
| 34 | 15:10 | 18:00 | KHH | SGN | Air | 110 |
| 35 | 18:10 | 21:00 | KHH | SGN | Air | 110 |

reliability also drops much when increasing the required transit time between different transport modes from $w^* \leq 20$ to $w^* \leq 25$, except from the case $w \leq 30$. Namely, the reliability decreases only 0.5% from 0.933458. In short, it is recommended to put much effort into shortening transit time between the different transport modes. However, if the acceptable reliability is no lower than 0.9, the MTN will not qualify only at two levels of transit times: $W \leq (45, 25)$ and $W \leq (180, 25)$. A suggestion in this situation for the travel agent is controlling the transit time between the different transport modes at $w \leq 20$ (ie. up to 5% of the time threshold).

5. Practical case

This sub-section introduces a practical MTN in Fig. 4, constructed by 35 routes and 8 stations. The reliability to transport from the source – Changhua (CHU) in Taiwan to the sink – Haiphong (HPH) in Vietnam within 200 USD and 8 hours will be analyzed with various transit

times at a range of demands. The relevant information and results are shown in Tables 3, 4, and 5.

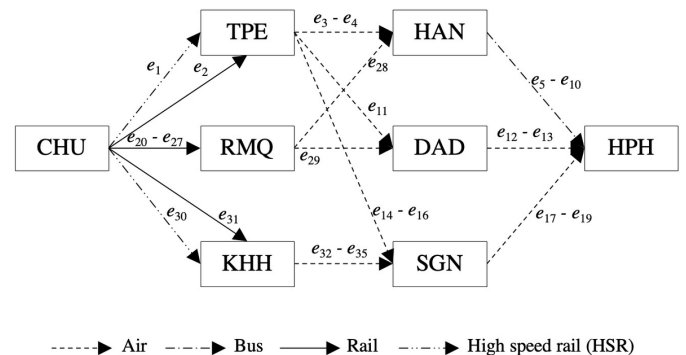


Fig. 4. A practical MTN

Table 4. The capacity probability

| Route (e_i) | Probability Pr (y_j passengers) | | | | | |
|-------------------|------------------------------------|--------|---------|---------|---------|---------|
| | 0 | 1 – 10 | 11 – 20 | 21 – 30 | 31 – 40 | 41 – 50 |
| 1, 8, 15, 22, 29 | 0.015 | 0.025 | 0.01 | 0.05 | 0.1 | 0.8 |
| 2, 9, 16, 23, 30 | 0.005 | 0.005 | 0.005 | 0.05 | 0.935 | |
| 3, 10, 17, 24, 31 | 0.01 | 0.05 | 0.02 | 0.06 | 0.05 | 0.81 |
| 4, 11, 18, 25, 32 | 0.015 | 0.05 | 0.055 | 0.88 | | |
| 5, 12, 19, 26, 33 | 0.02 | 0.005 | 0.015 | 0.03 | 0.93 | |
| 6, 13, 20, 27, 34 | 0.01 | 0.01 | 0.01 | 0.1 | 0.01 | 0.86 |
| 7, 14, 21, 28, 35 | 0.01 | 0.02 | 0.01 | 0.01 | 0.95 | |

Table 5. The reliability under time impacts of transit times and CPU time

| Demand (A) unit: passengers | 10 | 30 | 50 | 70 | 90 | 110 |
|-----------------------------|--------|--------|--------|--------|--------|--------|
| $W \leq (20, 40)$ | 0.9999 | 0.9997 | 0.996 | 0.98 | 0.8548 | 0.6593 |
| $W \leq (20, 45)$ | 0.9999 | 0.9997 | 0.9953 | 0.9763 | 0.8427 | 0.6176 |
| $W \leq (20, 55)$ | 0.9995 | 0.9933 | 0.9141 | 0.7319 | 0 | 0 |
| $W \leq (20, 115)$ | 0.9556 | 0.8755 | 0 | 0 | 0 | 0 |
| $W \leq (60, 40)$ | 0.9993 | 0.995 | 0.9391 | 0.8371 | 0 | 0 |
| $W \leq (60, 45)$ | 0.9991 | 0.9941 | 0.9323 | 0.8218 | 0 | 0 |
| $W \leq (60, 55)$ | 0.9652 | 0.9107 | 0 | 0 | 0 | 0 |
| $W \leq (60, 115)$ | 0.9556 | 0.8755 | 0 | 0 | 0 | 0 |
| CPU time (second) | 0.179 | 0.3013 | 0.7333 | 0.6634 | 0.3663 | 0.0967 |

According to Table 5, the impact of transit times is not significant at the demands $A = 10$ and $A = 30$ passengers, but that becomes more remarkable when the demand increases. In particular, the reliability is higher than 0.8 at the demand of 10 and 30 passengers, but it is zero at the higher demands with $W \leq (20, 115)$, $W \leq (60, 55)$, and $W \leq (60, 115)$. Besides, the reliability drops to zero at the demand from 90 passengers or more, except the transit times $W \leq (20, 40)$ and $W \leq (20, 45)$. From these results, we can conclude that the travel agent should keep the transit time no longer than 45 minutes between different modes to maintain the high reliability of transporting more than 30 passengers. To satisfy this requirement, the travel agent can arrange a private car or shuttle bus to carry passengers from the arrival gate of the previous route to the departure gate of the next route. To reliably transport up to 110 passengers, the required transit time should be lower than (20, 45). To keep the transit time between the same modes around twenty minutes, the travel agent should provide passengers the

stations' map with clear directions. Furthermore, the CPU time for the investigation is lower than one second for all experiments, which proves the efficiency of the proposed algorithm.

6. Conclusions

This study focuses on investigating the influence of transit time on the performance of a multistate transportation network reliability (MTN). The reliability, which is the probability to transport the given demand from the source to the sink within the time threshold and budget limitation, is the basis

of the investigation. This research proposes an algorithm that evaluates the MTN reliability under several levels of transit times. Instead of enumerating all possibilities, this study only takes significant levels of transit times under consideration. Namely, all examined transit time levels are generated through the maximal transit time vector of MTBPs. This paper also contributes a procedure determining all MTBPs and their maximal transit time vectors. The results of this study are used to provide some management suggestions such as the appropriate transit times to sustain reliability and transit time that exceeds the allowable reliability under certain circumstances. The proposed investigation is presented through a numerical example and a real case to demonstrate its application. In the future, estimating the impacts of other factors like economic or society [29] on reliability is a potential topic. Also, conducting the same analysis with a lack of probability information can earn more realistic findings.

References

1. Abdullah M, Ali N, Javid MA et al. Public transport versus solo travel mode choices during the COVID-19 pandemic: Self-reported evidence from a developing country. *Transportation Engineering* 2021; 5: 100078, <https://doi.org/10.1016/j.treng.2021.100078>.
2. Albalade D, Bel G. Tourism and urban public transport: Holding demand pressure under supply constraints. *Tourism Management* 2010; 31(3): 425-433, <https://doi.org/10.1016/j.tourman.2009.04.011>.
3. Albalade D, Fageda X. High speed rail and tourism: Empirical evidence from Spain. *Transportation Research Part A: Policy and Practice* 2016; 85: 174-185, <https://doi.org/10.1016/j.tr.2016.01.009>.
4. Bai G, Zuo MJ, Tian Z. Ordering Heuristics for Reliability Evaluation of Multistate Networks. *IEEE Transactions on Reliability* 2015; 64(3): 1015-1023, <https://doi.org/10.1109/TR.2015.2430491>.
5. Belobaba, Peter, Amedeo Odoni, and Cynthia Barnhart. *The global airline industry*. John Wiley & Sons, Ltd: 2015.
6. Boujelbene Y, Derbel A. The Performance Analysis of Public Transport Operators in Tunisia Using AHP Method. *Procedia Computer Science* 2015; 73: 498-508, <https://doi.org/10.1016/j.procs.2015.12.039>.
7. Choy KL, Sheng N, Lam HY et al. Assess the effects of different operations policies on warehousing reliability. *International Journal of Production Research* 2014; 52(3): 662-678, <https://doi.org/10.1080/00207543.2013.827807>.
8. Das S, Boruah A, Banerjee A et al. Impact of COVID-19: A radical modal shift from public to private transport mode. *Transport Policy* 2021; 109: 1-11, <https://doi.org/10.1016/j.tranpol.2021.05.005>.
9. Datta E, Goyal NK. Evaluation of stochastic flow networks susceptible to demand requirements between multiple sources and multiple

- destinations. *International Journal of System Assurance Engineering and Management* 2019; 10(5): 1302-1327, <https://doi.org/10.1007/s13198-019-00876-9>.
10. Habib A, Al-Seedy RO, Radwan T. Reliability evaluation of multi-state consecutive k-out-of-r-from-n: G system. *Applied Mathematical Modelling* 2007; 31(11): 2412-2423, <https://doi.org/10.1016/j.apm.2006.09.006>.
11. Huang C-F. Evaluation of system reliability for a stochastic delivery-flow distribution network with inventory. *Annals of Operations Research* 2019; 277(1): 33-45, <https://doi.org/10.1007/s10479-017-2600-6>.
12. IIAC. Incheon International Airport Corporation Annual Report 2016.
13. Jou R-C, Lam S-H, Hensher DA et al. The effect of service quality and price on international airline competition. *Transportation Research Part E: Logistics and Transportation Review* 2008; 44(4): 580-592, <https://doi.org/10.1016/j.tre.2007.05.004>.
14. Karimi B, Bashiri M. Designing a Multi-commodity multimodal splittable supply chain network by logistic hubs for intelligent manufacturing. *Procedia Manufacturing* 2018; 17: 1058-1064, <https://doi.org/10.1016/j.promfg.2018.10.080>.
15. Kim YK, Lee HR. Customer satisfaction using low cost carriers. *Tourism Management* 2011; 32(2): 235-243, <https://doi.org/10.1016/j.tourman.2009.12.008>.
16. Kong Z, Yeh EM. Correlated and cascading node failures in random geometric networks: A percolation view. 2012 Fourth International Conference on Ubiquitous and Future Networks (ICUFN), 2012: 520-525, <https://doi.org/10.1109/ICUFN.2012.6261764>.
17. Kos Koklic M, Kukar-Kinney M, Vegelj S. An investigation of customer satisfaction with low-cost and full-service airline companies. *Journal of Business Research* 2017; 80: 188-196, <https://doi.org/10.1016/j.jbusres.2017.05.015>.
18. Lee M-K, Yoo S-H. Using a Choice Experiment (CE) to Value the Attributes of Cruise Tourism. *Journal of Travel & Tourism Marketing* 2015; 32(4): 416-427, <https://doi.org/10.1080/10548408.2014.904259>.
19. Lesko S, Aleshkin A, Zhukov D. Reliability Analysis of the Air Transportation Network when Blocking Nodes and/or Connections Based on the Methods of Percolation Theory. *IOP Conference Series: Materials Science and Engineering* 2020; 714: 12016, <https://doi.org/10.1088/1757-899X/714/1/012016>.
20. Lin Y-K, Nguyen T-P, Yeng LC-L. Reliability evaluation of a stochastic multimodal transport network under time and budget considerations. *Annals of Operations Research* 2019, <https://doi.org/10.1007/s10479-019-03215-0>.
21. Liu X, Hou K, Jia H et al. The Impact-increment State Enumeration Method Based Component Level Resilience Indices of Transmission System. *Energy Procedia* 2019; 158: 4099-4103, <https://doi.org/10.1016/j.egypro.2019.01.825>.
22. Lunke EB. Commuters' satisfaction with public transport. *Journal of Transport & Health* 2020; 16: 100842, <https://doi.org/10.1016/j.jth.2020.100842>.
23. Maltese I, Zamparini L. Transport Modes and Tourism. *International Encyclopedia of Transportation*, Elsevier: 2021: 26-31, <https://doi.org/10.1016/B978-0-08-102671-7.10401-4>.
24. Mattrand C, Bourinet J-M. The cross-entropy method for reliability assessment of cracked structures subjected to random Markovian loads. *Reliability Engineering & System Safety* 2014; 123: 171-182, <https://doi.org/10.1016/j.res.2013.10.009>.
25. Nguyen T-P, Lin Y-K. Reliability assessment of a stochastic air transport network with late arrivals. *Computers & Industrial Engineering* 2021; 151: 106956, <https://doi.org/10.1016/j.cie.2020.106956>.
26. Niu Y-F, Xu X-Z. Reliability evaluation of multi-state systems under cost consideration. *Applied Mathematical Modelling* 2012; 36(9): 4261-4270, <https://doi.org/10.1016/j.apm.2011.11.055>.
27. O'Connell JF, Williams G. Passengers' perceptions of low cost airlines and full service carriers: A case study involving Ryanair, Aer Lingus, Air Asia and Malaysia Airlines. *Journal of Air Transport Management* 2005; 11(4): 259-272, <https://doi.org/10.1016/j.jairtraman.2005.01.007>.
28. Siu BWY, Lo HK. Doubly uncertain transportation network: Degradable capacity and stochastic demand. *European Journal of Operational Research* 2008; 191(1): 166-181, <https://doi.org/10.1016/j.ejor.2007.08.026>.
29. Wang T, Li B, Zhang G. Application of Panel Data Model to Economic Effects of High-Speed Railway. *International Journal of Performability Engineering* 2020; 16(7): 1130-1138, <https://doi.org/10.23940/ijpe.20.07.p15.11301138>.
30. Virkar AR, (India)PDM. A Review of Dimensions of Tourism Transport affecting Tourist Satisfaction. *Indian Journal of Commerce & Management Studies* 2018; IX(1): 72, <https://doi.org/10.18843/ijcms/v9i1/10>.
31. Yeh W-C. A simple algorithm for evaluating the k-out-of-n network reliability. *Reliability Engineering & System Safety* 2004; 83(1): 93-101, <https://doi.org/10.1016/j.res.2003.09.018>.
32. Yeh W-C. A new approach to evaluate reliability of multistate networks under the cost constraint. *Omega* 2005; 33(3): 203-209, <https://doi.org/10.1016/j.omega.2004.04.005>.
33. Chiu Y-H, Nguyen T-P, Lin Y-K. Network Reliability of a Stochastic Online-Food Delivery System with Space and Time Constraints. *International Journal of Performability Engineering*; 17(5): 433-443, <https://doi.org/10.23940/ijpe.21.05.p3.433443>.
34. Zahraee SM, Golroudbary SR, Shiwakoti N et al. Economic and environmental assessment of biomass supply chain for design of transportation modes: strategic and tactical decisions point of view. *Procedia CIRP* 2021; 100: 780-785, <https://doi.org/10.1016/j.procir.2021.05.044>.
35. Zuo MJ, Tian Z, Huang H-Z. An efficient method for reliability evaluation of multistate networks given all minimal path vectors. *IIE Transactions* 2007; 39(8): 811-817, <https://doi.org/10.1080/07408170601013653>.

An evaluation method of preventive renewal strategies of railway vehicles selected parts

Indexed by:



Jakub Lewandowski^a, Stanisław Młynarski^b, Robert Pilch^{a,*}, Maksymilian Smolnik^a, Jan Szybka^c, Grzegorz Wiązania^a

^aAGH University of Science and Technology, Faculty of Mechanical Engineering and Robotics, Al. A. Mickiewicza 30, 30-059 Kraków, Poland

^bCracow University of Technology, Faculty of Mechanical Engineering, ul. Warszawska 24, 31-155 Kraków, Poland

^cUniversity of Applied Sciences in Tarnów, Polytechnic Faculty, ul. Mickiewicza 8, 33-100 Tarnów, Poland

Highlights


- A method for evaluating the renewal strategies of wearing machine parts was developed.
- Calculations were made for a real case from railway operational practice.
- Obtained results indicate possibilities of improving the researched renewal strategy.
- The developed method is useful for services responsible for planning inspections.

Abstract

The aim of the work was to develop a method of verification of the preventive renewal strategies, which enables a simulation evaluation of the effects of the application of a specific schedule of inspections of parts that are important in the operation of complex renewable technical objects. Using it requires having an already established schedule of inspections, and the result of applying the method is determined by indicators that assess the usefulness of the strategy, even before implementation.

The developed computational tool was used to evaluate the renewal strategy of the current collector contact plates. Based on the real operational data, several renewal intervals were considered, determining the frequency of events involving the plate covering a specific mileage, from exceeding the wear control limit value to the next inspection (replacement). The proposed verification method is an important tool for testing and planning technical inspections for systems and elements with planned wear, and parts are periodically replaced.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

preventive renewal, rail vehicles, renewal strategy, contact plates.

1. Introduction

The failures of technical objects that occur in the process of their use have various causes. In particular, failures caused by random factors and degradation processes of parts of machinery [8] (e.g. wear, fatigue, corrosion, erosion, ageing) can be distinguished, cf. [10, 11]. These processes start with the beginning of an object's use, and for a long time proceed with no significant impact on the object's proper operation (that is performing the required functions). The object, however, stops working properly when the impact of the degradation processes exceeds a certain threshold state (see [9]) resulting from structural conditions (excessive clearance between the interacting parts, inadequate cross-sectional area or a change of parts' geometric dimensions, excessive surface roughness of interacting parts, too large proportion of corroded surface in the overall surface, etc.). In such cases in order to enable further correct functioning of the object it becomes indispensable to regenerate or replace its failed parts [20].

In planning the process of an object's use and maintenance the important problem that should be analysed is (apart from the nature of the events leading to failure) the effects of failures. If the failure involves a threat or considerable losses (e.g. to human life and health, environ-

mental contamination, interruption in services provision, a secondary failure of other parts or subassemblies of a machine resulting from a primary failure of another part), the aim of developing an object's use and maintenance strategy is to prevent failures of this type [23].

The negative effects of failures resulting from the degradation processes of parts of technical objects can be restricted in several ways: continuous verifying of the object's technical state (monitoring and corrective maintenance actions if necessary) or preventive replacement of parts (when they are still in the availability state) [13, 18]. The determination of the optimum time of preventive replacement can be assisted by relevant preventive maintenance models in which reliability characteristics most frequently estimated on the basis of the data on objects' failures history are used. Such models include block replacement strategy, age replacement strategy, etc. [28]. When the information on the technical condition (the extent of elements' wear) can be obtained while the object is working a condition-based maintenance strategy can be employed, which is sometimes more effective [17, 27].

Technical means subject to intensive wear during their use [2, 3], whose replacement is most frequently caused by their reaching a cer-

(*) Corresponding author.

E-mail addresses: J. Lewandowski - jlewando@agh.edu.pl, S. Młynarski - mlynarski_st@poczta.onet.pl, R. Pilch - pilch@agh.edu.pl, M. Smolnik - smolnik@agh.edu.pl, J. Szybka - szybja@agh.edu.pl, G. Wiązania - wiazania@agh.edu.pl

tain limit value of a given dimension (which is detected in a scheduled inspection) include current collector contact plates used in rail vehicles. Their correct operation determines the proper interaction of the vehicle with the transport infrastructure and proper electric power supply of the vehicle itself, cf. [15]. This translates into operational safety and continuity of the provision of transport services, which determines the economic results of carriers [25]. Problems of this kind are discussed in many studies [13].

In the present paper an original maintenance strategy verification method is proposed. It enables the simulation-based estimation of the effect of the adopted inspection strategy of parts most significant in the rail vehicle use (cf. [19]). It was assumed that in the framework of inspection procedures the current technical condition of tested current collector contact plates is estimated and (each time) the decision is taken as to their replacement or further operation (until the next scheduled inspection). It was also assumed that the given contact plate, degrading during operation, should be replaced on exceeding a limit control value of its wear (cf. [4]). This value is selected with a certain allowance so that the contact plate that has exceeded it owing to wear continues to operate correctly (for some time). Consequently, the assumed wear control limit value is not identical with the wear extent at which the contact plate can no longer perform its function (fails).

The proposed method is a tool assisting the technical services who schedule the inspection periods, helping them select preventive maintenance strategies that should deliver better results of the actual use of objects. This method is, then, not a typical maintenance scheduling model – understood as a mathematical model serving merely to determine the optimal periods of preventive replacement of objects. Its application requires a certain predefined inspection schedule, and its result is expressed with numerical indicators that provide an evaluation of the usability of the proposed strategy before it is implemented. These results can be referred to the evaluation of other (modified) strategies in order to select one that yields the most favourable effects or better fits in with maintenance project of the given object.

In practice also other strategies of current collectors use, facilitating their maintenance, but requiring the contact plates and current collectors of a certain design are employed. In one of these state-of-the-art technical solutions are applied which enable automatic taking the current collector out of operation on its exceeding the wear control limit specified in the design. In such emergency situation the safe solution is to drop the current collector to avoid its contact with the overhead line, which prevents damage to the interacting elements. This is done by a group of solutions called ADD (Automatic Drop Devices) by one of the leading manufacturers of current collectors – STEMMANN-TECHNIK [30]. The most popular ones are based on two technologies, mechanical and pneumatic. In the former one, when the current collector shoe rotation exceeds the allowable values (e.g. due to a collision with a broken element of the contact wire) the mechanism loosening the tension spring that holds the current collector in the top position is activated – the collector is dropped automatically [31]. The other solution is based on the pneumatic system in which the pressure that allows the current collector to be raised to the top position is maintained. When the pressure is inadequate, the current collector cannot be raised. A breach in the pneumatic circuit means pressure drop – in the matter of seconds the collector returns to the lower position. This can happen in the case of allowable limit wear of the graphite contact plate being reached – when the relevant friction device is damaged, the contact wire gets into contact with the compressed-air conduit in the current collector shoe, which becomes worn. When the pneumatic system is breached, the pressure is reduced and the current collector is dropped even before damage.

Owing to the application of the solutions described above the contact plate can be in operation to the value of the wear allowable limit, due to which it can be fully used over the entire durability and the moment of its failure is recorded accurately. The employment of this fact for inspections intervals optimisation will help make the maintenance

processes management more effective, which will result in both the financial result and the reduction of the number of failures as well as cases of using current collectors whose contact plates exceed the wear control limit values.

Although there are technical solutions that enable frequent, automated evaluation of the technical state of current collector contact plates [13] or taking out of operation the current collectors whose contact plates have been worn completely with no secondary, costly damage, it is still useful to predict and estimate the vehicle mileage to the moment of the collector contact plate reaching the limit value of wear, cf. [6]. Firstly, it enables planning in advance maintenance procedures connected with contact plates replacement, following which requires the provision of an adequate number of staff, service stands, tools and spare parts (new contact plates), cf. [14, 16]. Maintenance works planning, in turn, provides a basis of estimating the operating costs [1]. Secondly, when the presented design solutions of current collectors are not used in vehicles, it is indispensable to schedule their inspections adequately to the needs resulting from the durability of their parts so as to prevent their reaching the wear control limit while the vehicle is being used. Thirdly, if the solutions based on taking out of operation the current collectors at the moment their contact plates have been worn completely were employed and if the only maintenance strategy was their post-failure replacement, the reliability of vehicles would be affected negatively as they would be used regularly with a (at least one) unavailable current collector.

In view thereof, the design solutions based on degraded current collectors being automatically taken out of operation seem a valuable protection against serious consequences of uncontrolled wear limit being reached by a contact plate. The technical solutions of automated measurements of contact plates may definitely facilitate the scheduling of parts replacement over a short time horizon. However, the operation process course and costs planning over a longer time horizon requires the application of the tools of the theory of reliability and the renewal theory [24, 26].

2. Characteristics of the proposed calculation model

To enable the evaluation of the expected results of the application of the inspections schedule and renewal strategy a simulation calculation model has been developed.

Figure 1 presents a division of object's operation time horizon (T_H) into intervals. It is based on periodical inspections performed to detect whether the wear control limit has been exceeded, according to a defined schedule.

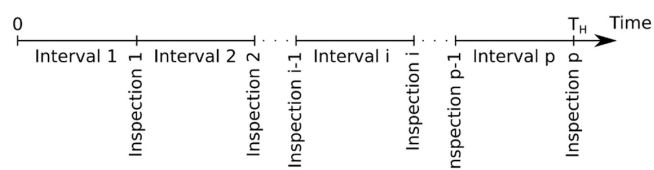


Fig. 1. Object's operation time horizon division into periodical inspections-based intervals

After exceeding the wear control limit the object continues to operate until the next periodical inspection during which it is detected. The time to inspection, shown in Figure 2, between exceeding the wear control limit and the next inspection, is particularly important because of inspection scheduling and determination of the wear control limit.

To estimate the time to inspection the simulation method, whose general description can be found in [21], was used. The graph of transitions for an object's reliability-operation states is shown in Figure 3, where the absorption state denotes the object's unavailability state.

An algorithm of a single iteration is shown in Figure 4, where object's operation time horizon in simulation was marked as T_H .

The proposed algorithm for obtaining simulation data has been employed for the evaluation of several renewal strategies of a selected technical object.

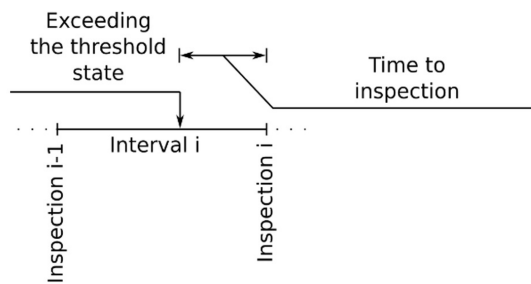


Fig. 2. Time to next inspection in the interval in which wear control limit has been exceeded

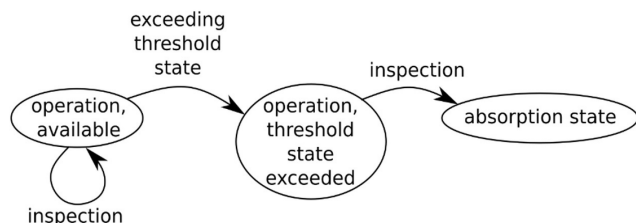


Fig. 3. Transitions of an object's states

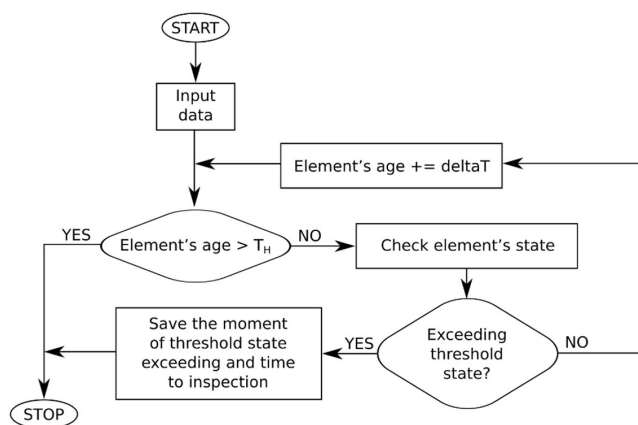


Fig. 4. Algorithm of simulation single iteration

3. Characteristics of the tested object and data on its failures

The study was performed for rail vehicle current collector contact plates, in which case the technical condition inspections are a basis for the decision of the object's replacement or continuation in the operation state until the next scheduled inspection. Contact plates belong to very important elements that guarantee the continuity of current flow between the overhead contact line and the vehicle [5, 6]. These elements, small in dimension, must meet the requirement of good electrical conductivity in a variety of atmospheric conditions while preserving low friction coefficient as the current is drawn with a vehicle in both standstill and in motion, not infrequently at a very high speeds [3, 12, 22].

Overhead contact lines are made of copper which has very good electrical properties and adequate resistance to both mechanical and climate induced failures. When contact plates are also made of copper high strength current can be drawn (over 1200 [A]), which is needed for the start-up of heavy trains. The possibility of high currents transfer is counterbalanced by the unfavourable properties of the contact joint of two elements made of the same materials – the static friction coefficient of such a joint is 1,5. It is reduced by the contamination of surfaces with oxides, but this phenomenon improves resistance. Owing to the above, the contact plates made of copper are subject to relatively high abrasive wear despite the fact that their design envis-

ages the application of lubricants between the contact plates mounted on the shoe [32].

In Poland contact plates made of graphite are much more commonly used. The material contains 85% carbon and the other components are copper and other additives. Carbon plates have very good friction coefficient which does not exceed 0,15 for the connection with the copper wire of the overhead contact line. The chemical composition of a given contact plate is modified according to its application – conduction of very high currents [2]. Moreover, carbon plates are a cheaper solution (mainly due to the limited percentage of expensive copper) [32].

In 2011 the Polish Railways (PKP Polskie Linie Kolejowe S.A.) introduced obligatory use of graphite contact plates on the lines under their control [25].

Regardless of the contact plate used, the nature of the current collector's operation inherently involves abrasive wear of the contact elements [7, 12, 29]. Exceeding the wear control limit of the contact plate may result in serious consequences – from degraded quality of current conduction, failure of the whole current collector, to breaking the overhead contact line. Therefore, very strict monitoring of this element is extremely important, because its timely replacement guarantees trouble-free operation and no need for costly and long-lasting failure.

The current collector can be monitored visually at any time. Then, any point damage that may lead to contact plate failure can be easily detected. The method of current collector's failures evaluation in the scope of employed devices and inspection intervals depends on the individual practice of the operator. It takes place regularly (e.g. every 3 000 [km]) and on its basis decision is taken as to the contact plate's replacement or its continued operation until the next inspection.

To estimate the probability distribution of the operation time-to-replacement of all of the selected current collector contact plates the data gathered in the operation and maintenance of several dozen electric locomotives series EU07, in which AKP 4E type current collectors are employed – shown in Figure 5, were used.

On the basis of these data tests of the goodness of fit of the operation time to wear-induced replacement (expressed in the mileage in kilometers) with theoretical probability distributions (Weibull, normal, exponential, gamma) were performed. For this purpose Statistica 13.1 software was used. After Kolmogorov-Smirnov test and χ^2 test, the best fit was obtained for three-parameter Weibull distribution. The tests were performed at the significance level of $\alpha = 0,05$. The results are given in Table 1 and Figure 6.

The probability density function and values of the estimated parameters of three-parameter Weibull distribution were adopted as formula (1):

$$f(t) = \alpha \left(\frac{1}{\beta} \right)^\alpha \cdot (t - \theta)^{\alpha-1} \cdot \exp \left(- \left(\frac{t - \theta}{\beta} \right)^\alpha \right); t > \theta \quad (1)$$

where:

- $\beta = 17,007 \cdot 10^3$ [km] – parameter of scale,
- $\alpha = 1,361$ – parameter of shape,
- $\theta = 10 \cdot 10^3$ [km] – parameter of shift.

In the data set under consideration, no object was replaced because of the dominant forms of wear of the contact plates, which resulted in their thickness decreasing before reaching the mileage of 10000 [km].

The inspections schedule and replacements strategy which in their basic version (applied for the sets of contact plates analysed in the presented study) lies in a periodic inspection of plates' technical condition and the decision taken on this basis as to their replacement, together with the adopted probability distribution of plates' time-to-replacement were used for the presentation of the calculation model proposed in the paper.

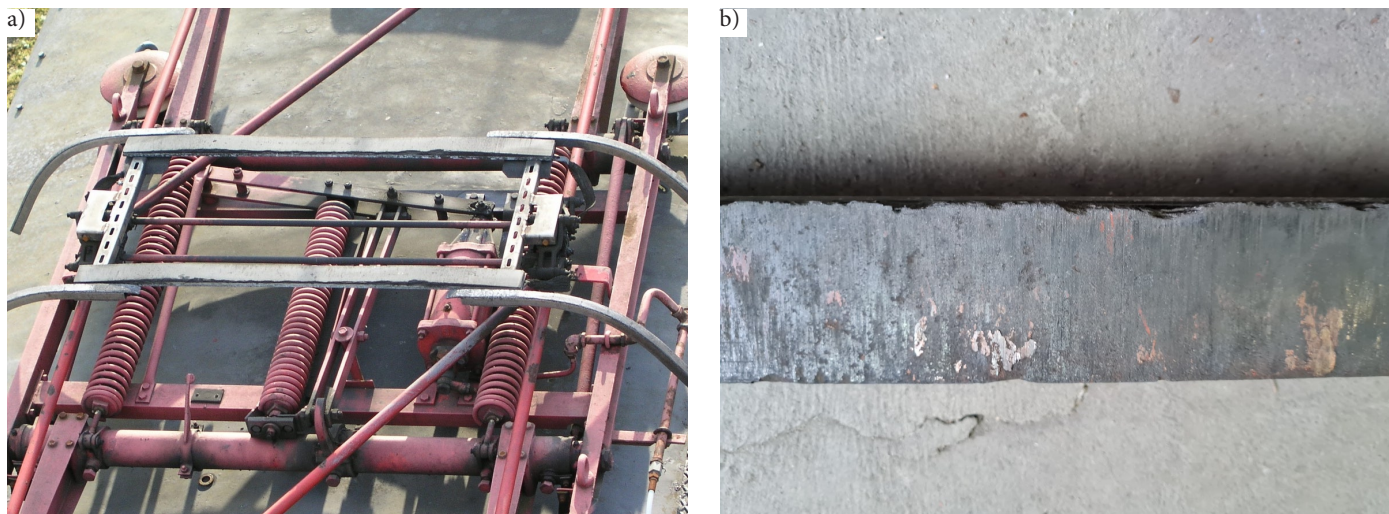


Fig. 5. AKP 4E: a) current collector; b) an example of a contact plate

Table 1. Results of test on data goodness of fit with Weibull shifted distribution

| Distribution | Characteristic value in test K-S | Test K-S p-value | Characteristic value in test χ^2 | Test χ^2 p-value |
|-------------------------|----------------------------------|------------------|---------------------------------------|-----------------------|
| three-parameter Weibull | 0,0825 | 0,980 | 0,483 | 0,785 |

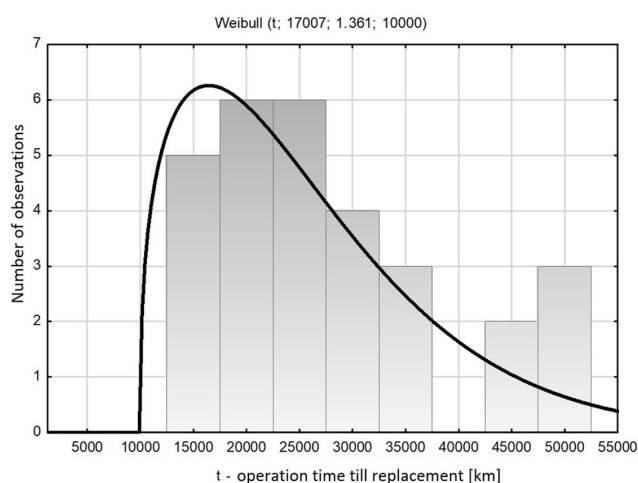


Fig. 6. Histogram of data and density function of three-parameter Weibull distribution

4. An example of an analysis of preventive renewal strategy

In the framework of verification of the developed calculation model simulation experiments were performed for the statistically determined probability distribution. This distribution reflects the operation-to-replacement mileage of a contact plate, which replacement was done based on the plate attaining or exceeding the adopted contact plate wear control limit. This means that this distribution was obtained only on the basis of the values of mileage which are a multiple of inspections interval. In the presented analysis, however, it is used as a model on the basis of which the potential moments of exceeding the wear control limit are identified by means of simulation. In the simulation experiment, as in real-life practice, the maintenance (replacement) as a response to this fact, can be undertaken only during the inspection, that is at one of the moments established in advance resulting from the inspection scheduling. However, in the experiment the potential mileage during which the wear control limit has been

exceeded is simulated accurately, which allows checking the delay-time between contact plate wear control limit being exceeded and the plate replacement. It is additionally assumed that in simulation the replacement of the contact plate whose thickness is close to the wear control limit, which sometimes occurs in the real operation, is never performed. It should be emphasized that in the simulation experiment the replacement can take place only after this value has been exceeded, in the forthcoming inspection. For this reason, all the contact plates used in simulation work longer (if only for a short while) than until the wear control limit is exceeded.

The calculations were performed for five different (3, 6, 12, 18 and 36 [thousand km]) fixed values of mileage after which the current collector scheduled inspection is done. In each experiment the moment of attaining the contact plate wear control limit defined with an accuracy of 100 [km], the simulation time horizon is 72 [thousand km], and the number of iterations 10 thousand. As assumed before, the contact plate is replaced only after its wear control limit has been exceeded, during the nearest inspection (so there are no typical preventive replacements). Another assumption was that this limit had been selected with a certain allowance so that its being exceeded does not involve any immediate interference in the object's proper functioning. The result of each iteration is the mileage that the given contact plate covers from the moment of attaining the limit wear to the nearest scheduled inspection (which equals its replacement).

Figures 7 – 11 illustrate the frequency of events of a contact plate covering a certain mileage, from exceeding the wear control limit until the forthcoming inspection (replacement). In each case these events are analyzed for subsequent mileage ranges of 1 [thousand km] preceding the inspection.

The obtained bar charts illustrate the dependence of the expected distribution of mileage (the usage time) after exceeding the wear control limit on the fixed inspections (interval) schedule. It is far from obvious when only the initial probability distribution is analysed. As can be seen, subsequent cases of exceeding the wear control limit add up to others within individual intervals preceding inspection.

The analysis indicates how long the given part of objects continues to be used after the wear control limit has been exceeded. It is useful in the evaluation of the analysed inspection schedule and renewal strategies because the proportion of objects that will be used over an excessive period of time can be identified. And it should be remembered that the wear control limit considered in this study is nominal in nature and is not equivalent to the wear extent which prevents proper functioning of an object. A settled wear control limit is therefore merely a supporting value, indispensable in taking the decision of an object's replacement. Since the level at which an object no longer op-

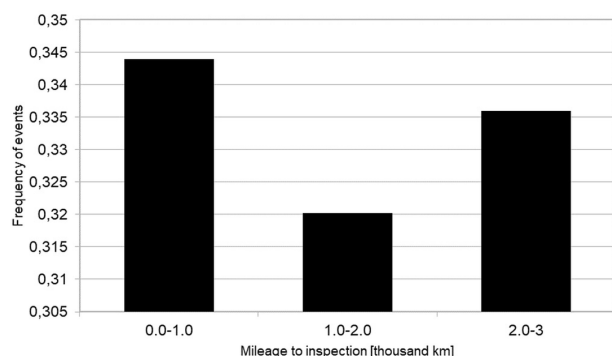


Fig. 7. Histogram of mileage from exceeding wear control limit to forthcoming inspection (replacement) for inspection intervals of 3 [thousand km]

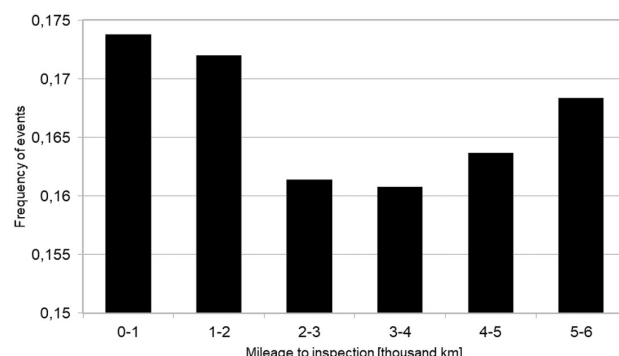


Fig. 8. Histogram of mileage from exceeding wear control limit to forthcoming inspection (replacement) for inspection intervals of 6 [thousand km]

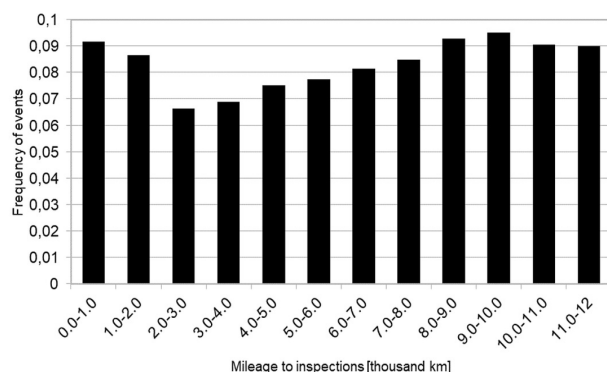


Fig. 9. Histogram of mileage from exceeding wear control limit to forthcoming inspection (replacement) for inspection intervals of 12 [thousand km]

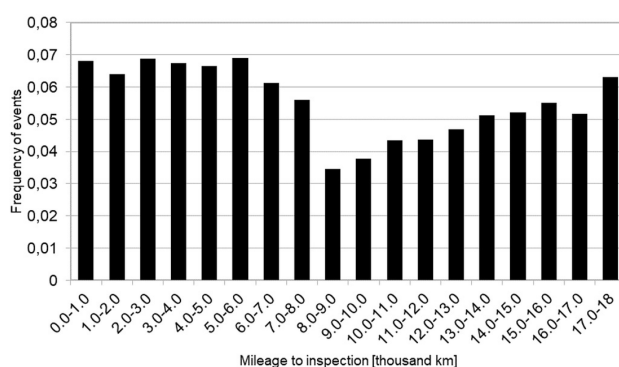


Fig. 10. Histogram of mileage from exceeding wear control limit to forthcoming inspection (replacement) for inspection intervals of 18 [thousand km]

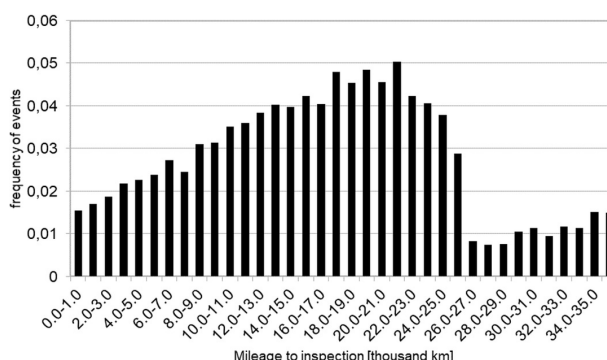


Fig. 11. Histogram of mileage from exceeding wear control limit to forthcoming inspection (replacement) for inspection intervals of 36 [thousand km]

erates correctly must be prevented, it should be replaced beforehand. The replacement should be performed at the mentioned settled wear control limit. A natural consequence of such an approach is that the selected wear control limit should guarantee the safe use of the object until the next inspection during which this value being exceeded will be identified. On the basis of the analysis of the presented results it can be stated what proportion of objects is used for a period longer than allowed by the selected allowance resulting from the adopted nominal allowable wear threshold, which is an estimate of a risk of the occurrence of a serious failure. An analysis of the presented results therefore indicates for how long (over what mileage) what number of contact plates is used after the adopted wear control limit has been exceeded, which constitutes an evaluation of the threat of the occurrence of a severe failure.

In the case of the use of the contact plates of AKP 4E current collectors in EU07 electrical locomotives, discussed in the present study, the probability of attaining the contact plate wear allowable limit can be established for each of the proposed inspection schedules.

According to the operational specifications, contact plate worn thickness g – from the nominal value to the wear control limit after which it is replaced as scheduled – is 12 [mm]. Between the wear control limit value and the wear allowable limit a thickness margin g_z of 5 [mm] is adopted. In accordance with the adopted probability distribution of contact plate wear, for the operational data specifying the contact plate wear control limit, the wear by the value of g occurs after the mileage of at least 10000 [km]. On this basis the contact plate maximum wear z_{max} per a mileage kilometer can be estimated:

$$z_{max} = \frac{g}{\theta} = \frac{12}{10000} = 1.2 \cdot 10^{-3} \left[\frac{\text{mm}}{\text{km}} \right] \quad (2)$$

On this basis an approximate estimation can be made (assuming the same wear mean rate) of the shortest mileage x_{min} after which the thickness margin g_z is used up if the wear control limit is attained between inspections:

$$x_{min} = \frac{g_z}{z_{max}} = \frac{5}{1.2 \cdot 10^{-3}} = 4167 [\text{km}] \quad (3)$$

This value enables the identification of the potential numerical proportion of working contact plates after reaching the wear allowable limit, that is, after the margin g_z is completely used up before the next inspection. This value can be specified for each inspection schedule analyzed in the presented simulations. The contact plates estimate numerical proportion indicates the probability that the contact plate will reach the wear allowable limit before the next inspection, which means that the current collector's proper operation will be disturbed. The probabilities in the proposed strategies are given in Table 2.

As can be noticed, the schedule with inspections every 3000 [km] mileage nearly ensures that the contact plate will not reach the wear

Table 2. Probability of contact plate reaching wear allowable limit in various inspection schedules determined with the use of the lowest mileage value till the margin g_z is completely used up

| Mileage between inspections in schedule [km] | 3000 | 6000 | 12000 | 18000 | 36000 |
|--|------|-------|-------|-------|-------|
| Probability of contact plate reaching wear allowable limit | 0 | 0,305 | 0,674 | 0,721 | 0,923 |

Table 3. Probability of contact plate reaching wear allowable limit in various inspection schedules determined with the use of the expected value of the mileage till the margin g_z is completely used up

| Mileage between inspections in schedule [km] | 3000 | 6000 | 12000 | 18000 | 36000 |
|--|------|------|-------|-------|-------|
| Probability of contact plate reaching wear allowable limit | 0 | 0 | 0,120 | 0,378 | 0,743 |

allowable limit between inspections, even in this pessimistic version which allows the fastest possible consumption of the g_z . This inspection schedule is employed by the operator of the contact plates analyzed in the study. In the other schedules the probability of contact plate reaching wear allowable limit is greater. Based on the results obtained, however, it can be stated that the zero probability of contact plate reaching wear allowable limit between inspections could also be attained by lengthening the period between inspections up to 4000 [km], that is, by about 30% compared with the schedule employed at present.

If, however, the expected value $E(T)$ of contact plate were after an adopted distribution was used as a basis, the analysis would be:

$$E(T) = \Gamma\left(1 + \frac{1}{\alpha}\right) \cdot \beta + \theta = 25573 \text{ [km]} \quad (4)$$

$$z_{sr} = \frac{g}{E(T)} = \frac{12}{25573} = 0.47 \cdot 10^{-3} \left[\frac{\text{mm}}{\text{km}} \right] \quad (5)$$

$$x_{sr} = \frac{g_z}{z_{sr}} = \frac{5}{0.47 \cdot 10^{-3}} = 10638 \text{ [km]} \quad (6)$$

For such a case, the probabilities of contact plate reaching wear allowable limit before the next inspection in the discussed strategies are shown in Table 3.

On the basis of the expected value of mileage x_{sr} after which the contact plate thickness margin g_z is used up, the schedule with inspections every 6000 [km], as the schedule with inspections every 3000 [km], gives a zero value of probability of contact plate reaching wear allowable limit between inspections. The results, however, should be interpreted remembering the fact that the margin g_z may be used up at a mileage smaller than would result from the expected value. Such an approach introduces a broader range of uncertainty in decision taking and increases the risk that might not be acceptable to the operator.

References

1. Ao Y, Zhang H, Wang C. Research of an integrated decision model for production scheduling and maintenance planning with economic objective. *Computers & Industrial Engineering* 2019; 137: 106092, <https://doi.org/10.1016/j.cie.2019.106092>.
2. Babyak M, Horobets V, Sychenko V, Horobets Y. Comparative tests of contact elements at current collectors in order to comprehensively assess their operational performance. *Eastern-European Journal of Enterprise Technologies* 2018; 6: 13–21, <https://doi.org/10.15587/1729-4061.2018.151751>.
3. Bucca G, Collina A. A procedure for the wear prediction of collector strip and contact wire in pantograph–catenary system. *Wear* 2009; 266: 46–59, <https://doi.org/10.1016/j.wear.2008.05.006>.
4. Cavalcante C A V, Lopes R S, Scarf P A. Inspection and replacement policy with a fixed periodic schedule. *Reliability Engineering & System Safety* 2021; 208: 107402, <https://doi.org/10.1016/j.ress.2020.107402>.
5. Chen G. Effect of the Staggering of a Contact Wire on Wear Behaviour of the Contact Strip with Electric Current. *Journal of Robotics and Mechanical Engineering Research* 2017; 2: 1–6, <https://doi.org/10.24218/jrmer.2017.21>.
6. Derosa S, Nàvik P, Collina A et al. A heuristic wear model for the contact strip and contact wire in pantograph – Catenary interaction for railway operations under 15 kV 16.67 Hz AC systems. *Wear* 2020; 456–457: 203401, <https://doi.org/10.1016/j.wear.2020.203401>.
7. Ding T, Chen G, Bu J, Zhang W. Effect of temperature and arc discharge on friction and wear behaviours of carbon strip/copper contact wire

5. Conclusions

The proposed algorithm enables an evaluation of the expected effects of inspections schedule and renewal strategy of given objects, and the analytic method defines a procedure applicable for the comparison of the effects of various strategies in order to select the most favourable one to apply in practice (without direct consideration of the cost of inspections and the

effects of failures).

In the studied case, the analysis of the results leads to a conclusion that lengthening the intervals between inspections by 1000 [km] (inspection every 4000 [km]) is safe. Such lengthening the intervals between inspections is justified economically and, moreover, offers the probability of contact plate reaching wear allowable limit acceptable to the operator. It should, however, be remembered that the intensity of contact plates' wear differs with the seasons of the year, which may be the subject of further, more detailed analyzes. The proposed method makes it possible to change the model used and conduct analyzes for various conditions and wear processes.

The potential of the proposed model can be further developed to include a differentiation of object's age-based inspection intervals, depending on the contact plate's mileage. Higher inspection rate prior to the expected wear control limit being reached enables restricting the working time after this value has been exceeded thus providing a basis for the reduction of the reserve (surplus) of the material of the degraded parts. It is conducive to more effective use of contact plates in the aspect of their actual operational durability.

To ensure the operational safety of the discussed types of current collectors, the solution proposed in the present study can be used successfully. And when the design allows, automation-based modern solutions of taking out of operation the collector with worn contact plate can be introduced additionally. The statistics-based forecasting models facilitate inspections scheduling and spare parts management, and the state-of-the-art diagnostic and design solutions help better use degrading parts thus protecting against failures resulting from, *inter alia*, the imperfections of forecasts.

The proposed method is an important tool for testing and planning of inspection schedules for systems and elements which are subjected to expected operational wear, and parts are replaced in a cyclic formula.

- in pantograph–catenary systems. *Wear* 2011; 271: 1629–1636, <https://doi.org/10.1016/j.wear.2010.12.031>.
8. Han X, Wang Z, Xie M et al. Remaining useful life prediction and predictive maintenance strategies for multi-state manufacturing systems considering functional dependence. *Reliability Engineering & System Safety* 2021; 210: 107560, <https://doi.org/10.1016/j.res.2021.107560>.
 9. Hu J, Chen P. Predictive maintenance of systems subject to hard failure based on proportional hazards model. *Reliability Engineering & System Safety* 2020; 196: 106707, <https://doi.org/10.1016/j.res.2019.106707>.
 10. Huang Q, Wu G, Li Z S. Design for Reliability Through Text Mining and Optimal Product Verification and Validation Planning. *IEEE Transactions on Reliability* 2021; 70(1): 231–247, <https://doi.org/10.1109/TR.2019.2938151>.
 11. Kang K, Subramaniam V. Integrated control policy of production and preventive maintenance for a deteriorating manufacturing system. *Computers & Industrial Engineering* 2018; 118: 266–277, <https://doi.org/10.1016/j.cie.2018.02.026>.
 12. Klapas D, Benson F A, Hackam R. Simulation of wear in overhead current collection systems. *Review of Scientific Instruments* 1985; 56(9): 1820–1828, <https://doi.org/10.1063/1.1138101>.
 13. Kordestani M, Saif M, Orchard M E et al. Failure Prognosis and Applications—A Survey of Recent Literature. *IEEE Transactions on Reliability* 2021; 70(2): 728–748, <https://doi.org/10.1109/TR.2019.2930195>.
 14. Lin B, Zhao Y. Synchronized Optimization of EMU Train Assignment and Second-level Preventive Maintenance Scheduling. *Reliability Engineering & System Safety* 2021; 107893, <https://doi.org/10.1016/j.res.2021.107893>.
 15. Lin S, Feng D, Sun X. Traction Power-Supply System Risk Assessment for High-Speed Railways Considering Train Timetable Effects. *IEEE Transactions on Reliability* 2019; 68(3): 810–818, <https://doi.org/10.1109/TR.2019.2896127>.
 16. Liu G, Chen S, Jin H, Liu S. Optimum opportunistic maintenance schedule incorporating delay time theory with imperfect maintenance. *Reliability Engineering & System Safety* 2021; 213: 107668, <https://doi.org/10.1016/j.res.2021.107668>.
 17. Liu X, Li J, Al-Khalifa K N et al. Condition-based maintenance for continuously monitored degrading systems with multiple failure modes. *IIE Transactions* 2013; 45(4): 422–435, <https://doi.org/10.1080/0740817X.2012.690930>.
 18. Mehmeti X, Mehmeti B, Sejdiu R. The equipment maintenance management in manufacturing enterprises. *IFAC-PapersOnLine* 2018; 51(30): 800–802, <https://doi.org/10.1016/j.ifacol.2018.11.192>.
 19. Mira L, Andrade A R, Gomes M C. Maintenance scheduling within rolling stock planning in railway operations under uncertain maintenance durations. *Journal of Rail Transport Planning & Management* 2020; 14: 100177, <https://doi.org/10.1016/j.jrtpm.2020.100177>.
 20. Młynarski S, Pilch R, Smolnik M et al. A Simulation Model for Regenerated Objects with Multiparameter Evaluation of Technical Condition Reliability Estimation. *Journal of KONBiN* 2019; 49: 7–30, <https://doi.org/10.2478/jok-2019-0023>.
 21. Młynarski S, Pilch R, Smolnik M et al. Simulation-Based Forecasting of the Reliability of Systems Consisting of Elements Described by a Number of Failure Probability Distributions. *Journal of KONBiN* 2020; 50: 63–82, <https://doi.org/10.2478/jok-2020-0028>.
 22. Nāvīk P, Derosa S, Rönquist A. On the use of experimental modal analysis for system identification of a railway pantograph. *International Journal of Rail Transportation* 2021; 9(2): 132–143, <https://doi.org/10.1080/23248378.2020.1786743>.
 23. Pricopie A, Frangu L, Miron M, Caraman S. An improved degradation model for preventive maintenance. 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), 2020: 483–488, <https://doi.org/10.1109/ICSTCC50638.2020.9259687>.
 24. Selech J, Andrzejczak K. Identification of Reliability Models for Non-repairable Railway Component: Selected Papers from the 18th International Conference on Reliability and Statistics in Transportation and Communication, RelStat'18, 17-20 October 2018, Riga, Latvia. *Lecture Notes in Networks and Systems*, 2019: 507–518, https://doi.org/10.1007/978-3-030-12450-2_49.
 25. Sitarz M, Hełka A, Mańka A, Adamiec A. Testing of Railway Pantograph. *Archives of Transport* 2013; 25–26(1–2): 85–95.
 26. Świdorski A, Borucka A, Grzelak M, Gil L. Evaluation of Machinery Readiness Using Semi-Markov Processes. *Applied Sciences* 2020. doi:10.3390/app10041541, <https://doi.org/10.3390/app10041541>.
 27. Vališ D, Žák L, Pokora O, Lánský P. Perspective analysis outcomes of selected tribodiagnostic data used as input for condition based maintenance. *Reliability Engineering & System Safety* 2016; 145: 231–242, <https://doi.org/10.1016/j.res.2015.07.026>.
 28. Werbińska-Wojciechowska S. Preventive Maintenance Models for Technical Systems. *Technical System Maintenance: Delay-Time-Based Modelling*, Cham, Springer International Publishing: 2019: 21–100, https://doi.org/10.1007/978-3-030-10788-8_2.
 29. Yang H, Hu B, Liu Y et al. Influence of reciprocating distance on the delamination wear of the carbon strip in pantograph–catenary system at high sliding-speed with strong electrical current. *Engineering Failure Analysis* 2019; 104: 887–897, <https://doi.org/10.1016/j.engfailanal.2019.06.060>.
 30. [<https://www.wabtec.com/uploads/outlinedrawings/Stemmann-Technik-brochure-Railway-Technology-Systems-English-Survey.pdf>] (accessed 03.2020).
 31. [<http://www.karma.se/uploads/1/8/4/1/18413739/add-pdf.pdf>] (accessed 03.2020).
 32. [<http://www.mitel.uz.zgora.pl/CD/2010/s347.pdf>] (accessed 03.2020).

Remaining useful life prediction of bearings with different failure types based on multi-feature and deep convolution transfer learning

Indexed by:



Chenchen Wu^{a,b}, Hongchun Sun^{a,b,*}, Senmiao Lin^{a,b}, Sheng Gao^{a,b}

^aSchool of Mechanical Engineering and Automation, Northeastern University, Shenyang 110819, China

^bKey Laboratory of Vibration and Control of Aero-Propulsion Systems of Ministry of Education, Northeastern University, Shenyang 110819, China


Highlights

- Spatial pyramid pooling extracts multi-scale degradation features of bearings.
- TL solves the inconsistent distribution of degraded data for different failed bearings.
- The SPP-CNN model shows a better prediction effect on the RUL of the bearing.

Abstract

The accurate prediction of the remaining useful life (RUL) of rolling bearings is of immense importance in ensuring the safe and smooth operation of machinery and equipment. Although the prediction accuracy has been improved by a predictive model based on deep learning, it is still limited in engineering because lots of models use single-scale features to predict and assume that the degradation data of each bearing has a consistent distribution. In this paper, A deep convolutional migration network based on spatial pyramid pooling (SPP-CNN) is proposed to obtain higher prediction accuracy with self-extraction of multi-feature from the original vibrating signal. And to consider the differences of the data distribution in different failure types, transfer learning (TL) added with maximum mean difference (MMD) measurement function is used in the RUL prediction part. Finally, the data of IEEE PHM 2012 Challenge is used for verification, and the results show that the method in this paper has high prediction accuracy.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

rolling bearings, Remaining useful life (RUL), Convolutional Neural Networks (CNN), Transfer learning (TL).

1. Introduction

As one of the most important components in rotating machinery, rolling bearings play a vital role in the safe operation of mechanical equipment [5]. According to relevant statistics, 45% to 55% of the failure cases of rotating machinery are caused by the failure of rolling bearings [19]. Accurate RUL prediction technology can ensure both the safety of operator and equipment in good condition, and it is of a certain significance for the predictive maintenance.

The current methods used to predict RUL can be summarized into four categories [12]: physical model-based methods [11], statistical model-based methods [29], artificial intelligence-based methods [22], and hybrid methods [26]. The physical model-based methods describe the degradation process of machinery through the failure mechanism of mechanical equipment and mathematical model. Although this method can theoretically explain the degradation state of machinery, as the complexity of the mechanical system becomes higher and higher, it is difficult to establish an ideal degradation model. These statistical model-based methods can achieve predictions under different working conditions, but it is usually assumed that the degraded signal follows a parameterized process model, which may not be the case in reality [33]. The data-driven method gets rid of the shackles of traditional methods, and the degraded state of the bearing can be

described based on the obtained bearing operating data. Therefore, the data-driven-based forecasting methods get wide attention. Recently, common models of data-driven methods gain very good effectiveness, such as Artificial Neural Network (ANN) [1], Support Vector Machine (SVM) [20, 24], Extreme Learning Machine (ELM) [28], etc. But each of these models is a shallow neural network that is of bad extraction ability and it is unable to directly mine the degraded information from the original data.

As a branch of machine learning, in recent years, deep learning emerges for its powerful feature extraction ability. Great progress has been made in image recognition, target detection, medicine, and other fields [13, 22, 23]. At present, the commonly used deep learning models in the mechanical field include Long Short-Term Network (LSTM) [30], Convolution Neural Network (CNN) [32], Stacked Denoising Autoencoding (SDA) [31], and Deep Belief Network (DBN) [21]. For instance, Wang et al [25] recurrent convolution layers were constructed to simulated the temporal correlation between different degradation states, and the variational inference was combined to measure the uncertainty of RUL prediction. It indicates that this neural network is obviously superior to other methods in terms of RUL prediction accuracy and convergence. Hinch et al [9] use the convolutional layer to extract the features from the original data, and the

(*) Corresponding author.

E-mail addresses: C. Wu - 2079879663@qq.com, H. Sun - hchsun@mail.neu.edu.cn, S. Lin - 490230707@qq.com, S. Gao - wsgs1415926@gmail.com

degradation process is captured by the LSTM layer to predict RUL. Wang et al [27] Transform original one-dimensional signal into the grey-scale image and use 2D-CNN network for feature extraction. Then the double Gaussian model is used to fit and predict the degradation curve. The results indicate that the method can predict the RUL of bearing, and this measurement has pretty good accuracy. Compared with the shallow neural network, the mentioned deep learning model has made some progress in the field of bearing RUL prediction., but two issues remain as follows:

1. Only the last layer feature is taken for the prediction of bearing RUL in most of the literature. Because the last feature is the most abstract feature, which makes the generalization ability of the network model worse, thus, the forecasting results of bearing RUL under various failure types cannot be accurate enough.
2. The impact of inconsistent bearing data distribution on the deep learning prediction model is not considered. Because the traditional deep learning model is suitable for the situation where the data distribution of the training set and the test set are consistent, however, even under the same working conditions and the same type of bearings, each bearing will show inconsistent degradation trends during the full life test of the bearing, resulting in bearing data that does not meet the assumptions of deep learning applications.

As a new learning paradigm in machine learning, transfer learning broadens the applicable conditions of deep learning. At present, it has been applied in the field of reliability. For example, Guo L et al [8] proposed a domain adaptive module to solve the difference between different bearing data distributions so as to realize bearing fault diagnosis across experimental platforms. Dong S et al [6] proposed a bearing degradation assessment model based on transfer learning and deep hierarchical feature extraction. Experiments show that the model can accurately identify the degraded stage of the bearing. Zhu J et al [33] applied the domain adaptive module proposed in Literature 23 to the field of bearing RUL prediction and successfully realized bearing RUL prediction under different working conditions. It can be seen that most applications of transfer learning in the mechanical field are dedicated to solving classification problems [6, 14], while regression problems have not been widely used [18]. However, transfer learning has great potential for simple prediction regression problems [15].

Therefore, in order to solve the above problems, a framework for RUL prediction of bearings based on SPP-CNNLTL is proposed. First, the degradation stage of the bearing is divided by a binary classification network. This method avoids human error caused by manual threshold division. Then, for the data in the degradation stage of the bearing, the frequency spectrum is extracted as input, and one-dimensional CNN is used as the feature extraction network. The SPP layer is used as the last pooling layer of CNN to achieve convolutional features observed from different directions. In addition, transfer learning based on the MMD function is introduced in the CNN model to solve the problem of low prediction accuracy caused by inconsistent bearing data distribution of different fault types. Finally, the method in this paper is verified by the IEEE PHM 2012 data set, and the results show that the prediction accuracy of bearing RUL is better than other models.

The contributions of this article are summarized as follows:

1. The spatial pyramid pooling layer is used to realize multi-scale feature extraction of input data, avoiding the shortcomings of insufficient bearing degradation information extracted.
2. Transfer learning is used to solve the problem of inconsistent distribution of bearing degradation data and failure data, so as to realize the deep learning model to predict the RUL of different failed bearings.
3. Propose an end-to-end prediction framework applicable to different faulty bearings, and promote the development of predictive maintenance technology for bearings.

The remainder of this paper is organized as follows: Section 2 describes the framework of the bearing remaining life prediction method proposed in this paper. The related theories of CNN and transfer learning networks are introduced, and the framework of the SPP-CNNLTL neural network is proposed. In Section 3, the experimental analysis based on the full life data set of the bearing shows the effectiveness of the method. The comparison with other model methods highlights the superiority of this method. Finally, conclusions are given, and some future research directions are proposed in Section 4.

2. Proposed framework

2.1. Overall overview

In engineering applications, due to bearing processing and manufacturing errors, assembly errors, and material defects of the bearing itself, the entire degradation process of the bearing from the initial use to the final failure shows different trends. This leads to the problem of differences in the data distribution between the degradation data of each bearing. This violates the assumption that deep learning requires the training set and test set to have the same data distribution, so it reduces the RUL prediction accuracy. Therefore, this paper proposes a framework for predicting the remaining life of bearings based on a multi-scale convolutional transfer learning model. The flow of the framework is shown in Figure 1. It can be seen from Figure 1 that the method in this article is mainly divided into two parts: the first part is the degradation stage division. This part uses the normal stage data and the severe stage data of the bearing to construct a data set, trains the two-class neural network and realizes the degradation stage Automatic division. This method avoids the human error caused by the trouble of manually setting the fault threshold in the traditional method and makes the recognition effect more objective. When the bearing enters the degradation stage, the second part starts to predict the RUL of the bearing based on the SPP-CNNLTL model. The model adds an SPP pool to solve the problem of the poor generalization ability of single-scale input. The domain adaptation technology in transfer learning is used to measure the difference between degraded data distributions in different directions and use the difference as a constraint condition of the prediction model so that the network model can learn the invariance between different failed bearing data.

2.2. Transfer learning

As a branch of machine learning, transfer learning can transfer learned knowledge in a different area, and its main idea is to find similarities between different datasets. Two basic concepts are mainly included in transfer learning, which are domain and task. The domain is the subject of learning, which is mainly composed of data and the probability distribution which can generate these data; Task is the goal of learning, which is mainly composed of tag and tag's corresponding function group. Thus, transfer learning can be expressed as follows: a labeled source domain $D_s = \{x_i, y_i\}_{i=1}^n$ and an unlabeled target domain $D_t = \{x_i\}_{i=1}^n$. They have different data distribution, $P_s(X_s) \neq P_t(X_t)$. The goal of transfer learning is to use labeled data D_s to learn the knowledge of the target domain D_t .

Domain adaptation is one of the research contents of transfer learning, which focuses on solving the problem of consistent feature space, consistent category space, and only inconsistent feature distribution. Domain adaptation mainly includes two strategies: One is to introduce the measurement function, minimizing its value to make the source domain and target domain obey the same distribution. Some measurement functions, such as Maximum Mean Discrepancy (MMD), KL divergence and CORAL, are often used. The other is to draw on the experience of the strategy of Generative Adversarial Network (GAN) --- adding domain classification module [4, 33].

Domain adaptive technology is proposed to solve the problem of different failure types of bearing RUL prediction, because domain

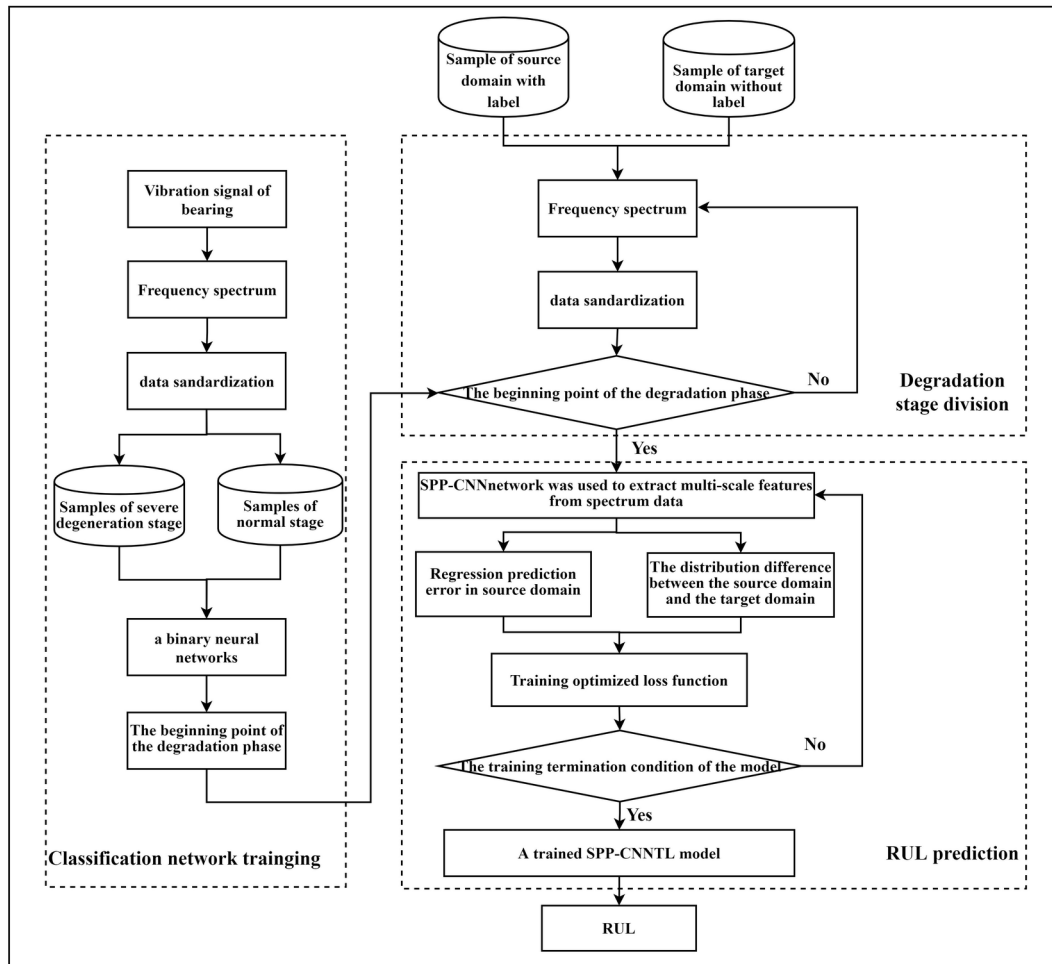


Fig. 1. Flow chart of the method proposed in this article

adaptive technology can perform classification and prediction when the data distribution of the training set and the test set are similar. Questions in this article is described in transfer learning language as follows:

1. To get some labeled degenerative data and to be used as training set, $D_s = \{\chi_s, P_s(X)\}$ and get some unlabeled degenerative data as test set, $D_t = \{\chi_t, P_t(X)\}$.
2. Assuming the feature space of the source of domain and the target domain is the same, $\chi_s = \chi_t$. But the marginal distribution of two domains is different, $P_s(X_s) \neq P_t(X_t)$.
3. A classifier $f: x_t \rightarrow y_t$ is adopted to improve the accuracy of prediction by using the auxiliary data that are composed of labelled data- D_s and partial unlabeled data- D_t .

2.3. CNN

CNN is a kind of feedforward neural network, which was first proposed by LeCun in 1989 and used for image processing [10]. The CNN network mainly consists of convolution layers, pooling layers, and full connection layers. The convolutional layer reduces the parameter amount of the model by capturing the local regional connection feature of input information and applying the weight sharing principle, and further reduces the amount of training data by combing the similar features through the pooling layer. In order to extract features from the data, the CNN model usually alternately stacks convolutional layers and pooling layers, and configures the output layer as a fully connected layer.

1. Convolutional layer

The convolution layer consists of a set of convolution kernels, which are the core of feature extraction. The convolution kernel

performs a convolution operation on the feature map output by the previous layer to achieve feature extraction of the local area. In addition, the convolutional layer also has the characteristics of weight distribution, which greatly reduces network parameters and avoids over-fitting. The specific convolutional layer operation is shown in the formula (1):

$$x_c^l = \sigma \left(\sum_{i=1}^{c^{l-1}} W_{i,c}^l * x_i^{l-1} + b_c^l \right) \quad (1)$$

where x_i^{l-1} is the output of channel i of $l-1$ layer, $W_{i,c}^l$ is the convolution kernel for layer l , b_c^l is bias, $*$ is convolution operation, x_c^l is the output of channel c of layer l . $\sigma(\cdot)$ is the activation function. In this paper, the ReLU function is used as the activation function of the CNN network because it has the ability to accelerate the convergence and alleviate the vanishing gradient problem. The calculation is as follows:

$$ReLU(x) = \max(0, x) \quad (2)$$

2. Pooling layer

The main purpose of the pooling layer is to reduce the parameters of the neural network. It is usually added between two convolutional layers, and the input of the convolutional layer at a specific connection position is summarized in the form of non-linear sampling to improve the computational efficiency of the network and keep the feature translation unchanged. Common pooling layers include aver-

age pooling, maximum pooling, etc. And maximum pooling is used in this paper partially. The equation (3) is as follows:

$$p_c^l = \max \left\{ x_{c \times k:(c+1) \times k}^l \right\} \tag{3}$$

where k is the length of pooling, p_c^l is the output of channel c layerl.

3. Spatial pyramid pooling

In order to solve the problem of inconsistent input image size, a spatial pyramid pool for target detection task is first proposed. SPP can extract features of different dimensions from the feature map by using pool kernels of various sizes, and stitch them to obtain multi-dimensional features. Therefore, this article adds SPP to the last layer of the CNN network model for multi-feature extraction to improve the generalization of the network.

4. Fully connected layer

The purpose of the fully connected layer is to perform regression or prediction tasks on the extracted features. After executing the SPP-CNN model in this article, the network will output multiple feature values and then pave them. The mapping between features and bearing RUL uses fully connected layers. The calculation formula (4) between complete connections is as follows:

$$h^l = \sigma^l \left(\left(W^l \right)^T \times v^{l-1} + b^l \right) \tag{4}$$

where σ^l is the activation function of the layer l, v^{l-1} is the output vector of layer l-1, W^l is the connection weight of the neurons in the l-th layer and the neurons in the l-1th layer, b^l is the bias, h^l is the output feature of the l-th hidden layer. The activation function of the output layer is the SoftMax function, and the other layers are the ReLU function.

2.4. SPP-CNNTL Learning model

The Figure 2 shows the framework of the SPP-CNNTL network model proposed in this paper. The network model mainly includes three parts: Multi-scale feature extraction module, regression prediction module, domain adaptive module. Among them, multi-scale feature extraction mainly uses the SPP-CNN model for feature extraction. The features that can represent bearing degradation information are extracted layer by layer by convolution and pooling operations from the input source domain and target domain. The regression prediction module is to predict the RUL of the bearing. The module uses the extracted multi-scale features as the judgment basis, and realizes the RUL prediction of the source domain samples through the fully

connected layer. The domain adaptation module is based on the data distribution difference between the source domain and the target domain in the specified layer, and uses the MMD function value as a measure to constrain the RUL prediction part to minimize the difference between the data distribution. The specific network model structure is shown in the Table 1.

Table 1. SPP-CNNTL Network Model diagram

| Layer | Module | Symbol | Operation | Parameter |
|-------|--------------------|---------|-----------------|-----------|
| 1 | Feature extraction | Input | Input signal | 1×2048 |
| 2 | | C1 | Convolution | 5×1×3 |
| 3 | | P1 | Pooling | 2 |
| 4 | | C2 | Convolution | 5×3×6 |
| 5 | | P2 | Pooling | 2 |
| 6 | | SPP | Multi-Pooling | / |
| 7 | | Flatten | / | 126 |
| 8 | Domain adaptive | FC1 | Fully-connected | 50 |
| 9 | | FC2 | Fully-connected | 10 |
| 10 | RUL prediction | FO | Sigmoid | / |

2.4.1. Domain adaptive model

Domain adaptive model is mainly to describe the difference among the data distribution of data set in some measures. Maximum mean difference is taken as the measurement function in this paper. This method measures the distance between two reproducing Hilbert space, which is a kernel learning method. The equation (5) is as follows:

$$MMD(h^s, h^t) = \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \phi(h_i^s) - \frac{1}{n^t} \sum_{i=1}^{n^t} \phi(h_j^t) \right\|_H^2 \tag{5}$$

where n^s the number of samples from the source domain, n^t is the number of samples from the target domain, $\phi(\cdot)$ is mapping which maps the original variable to the regenerative nuclear Hilbert space, $\|\cdot\|_H$ is the regenerative nuclear Hilbert space.

2.4.2. Target of optimization

The loss function of the proposed method are two parts:

1. Root mean square error term of the minimized regression task.

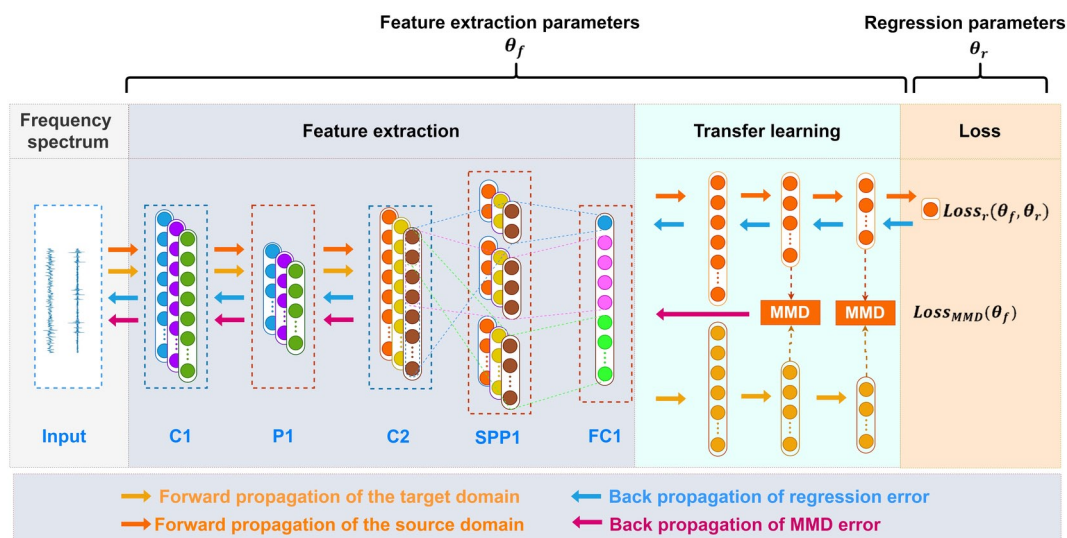


Fig. 2. SPP-CNNTL Network Model diagram

2. Minimized MMD term between the source domain and the target domain.

Loss function 1: The accuracy of RUL prediction of bearing is improved by minimizing differences in values. In other words, the main loss function is the difference between the predicted value and real labelled value. For regression tasks, the Mean Square Error (MSE) is the most commonly used as loss function. The equation is as follows:

$$Loss_r = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \quad (6)$$

where m is the size of batch of training set, y_i is the real label, \hat{y}_i is the label of prediction.

Loss function 2: The migration of the last two layers is selected after analysis: for the RUL prediction of bearing after the full connection layer, the difference among different domains is minimized after MMD is added into different layers. The equation is as follows:

$$Loss_{MMD1} = \frac{1}{m_s^2} \sum_{i=1}^{m_s} \sum_{j=1}^{m_s} k(f1_i^s, f1_j^s) + \frac{1}{m_t^2} \sum_{i=1}^{m_t} \sum_{j=1}^{m_t} k(f1_i^t, f1_j^t) - \frac{1}{m_s m_t} \sum_{i=1}^{m_s} \sum_{j=1}^{m_t} k(f1_i^s, f1_j^t)$$

$$Loss_{MMD2} = \frac{1}{m_s^2} \sum_{i=1}^{m_s} \sum_{j=1}^{m_s} k(f2_i^s, f2_j^s) + \frac{1}{m_t^2} \sum_{i=1}^{m_t} \sum_{j=1}^{m_t} k(f2_i^t, f2_j^t) - \frac{1}{m_s m_t} \sum_{i=1}^{m_s} \sum_{j=1}^{m_t} k(f2_i^s, f2_j^t)$$

$$Loss_{MMD} = Loss_{MMD1} + Loss_{MMD2} \quad (7)$$

where $Loss_{MMD1}$ is the value of the last layer, $Loss_{MMD2}$ is the inverted second layer, $k(\cdot)$ is the kernel function, m_s is the number of source domain samples, m_t is the number of the target domain samples.

The final total loss function is as follows:

$$Loss = Loss_r + \lambda Loss_{MMD} \quad (8)$$

where hyperparameter λ decide the effect of MMD differences on prediction.

And set the parameter of feature extractor as θ_f , and set the parameter of regression prediction of bearing RUL as θ_r . The equation 8 can be rewritten as follows:

$$Loss(\theta_f, \theta_r) = Loss_r(\theta_f, \theta_r) + \lambda Loss_{MMD}(\theta_f) \quad (9)$$

Adam optimizer is used to minimize the loss function and to find the saddle point of the loss function. The equation is as follows:

$$\begin{aligned} \theta_f &\leftarrow \theta_f - \eta \left(\frac{\partial Loss_r}{\partial \theta_f} + \lambda \frac{\partial Loss_{mmd}}{\partial \theta_f} \right) \\ \theta_r &\leftarrow \theta_r - \eta \left(\frac{\partial Loss_r}{\partial \theta_r} \right) \end{aligned} \quad (10)$$

where η is learning rate.

3. Application of the proposed method

3.1. Introduction of data set

IEEE PHM 2012 Challenge [16] is adopted to verify the effectiveness of the method proposed in this paper. Experiment platform of PRONSTIA is constructed as the Figure 3. The test-bed consists of two parts: part of experimental simulation and part of measurement. The power of the experimental simulation is output by a motor with a power of 250 W. And the load simulation is applied to the bearing to accelerate the degradation of the bearing by applying a radial force load. The measurement portion adopts an acceleration sensor whose sampling frequency is 25.6 kHz and the acquisition channel is two channels in the horizontal and vertical direction. A signal sample is collected every 10s, and the length of the collected time is 0.1 s.

The data set contains bearing work data under three different loads. Working-condition 1: under 1800 rpm and 4000 N; Working-condition 2: 1650 rpm and 4200 N; Working-condition 3: 1500 rpm, 5000 N. Total 17 data sets of bearing are acquired which are working to failure. In condition 1, there are 7 bearings numbered from 1-1 to 1-7; In condition 2, there are 7 bearings numbered from 2-1 to 2-7; In condition 3, there are 3 bearings numbered from 3-1 to 3-3. This paper selects the bearing in condition 1 for testing, and its partial degradation data is shown in Figure 4. Although the bearings are in the same working condition, they behave differently in degradation process. As pointed out in literature 3, under the working-condition 1, the bearings, 1-1 1-3 1-4, belongs to the same type of progressive degradation failure; the bearings, 1-2 1-5 1-6 1-7, belongs to the same type of sudden burst degenerate failure.

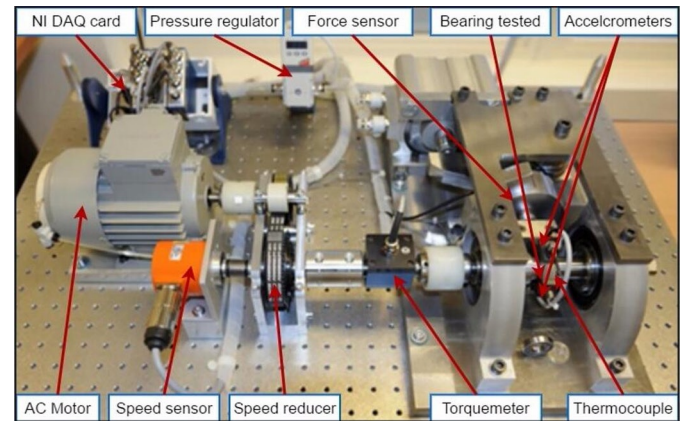


Fig. 3. The experimental platform

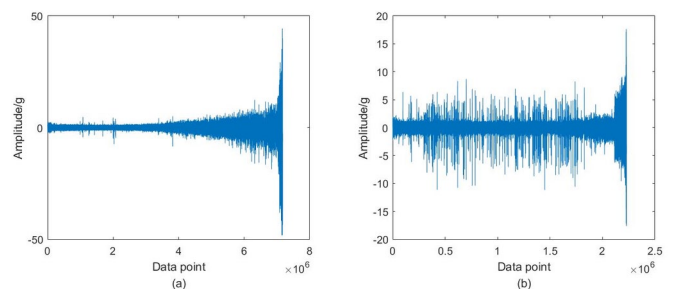


Fig. 4. Bearing degradation data under working-condition 1, (a) bearing 1-1; (b) bearing 1-2.

3.2. Starting point identification of degradation stage

The bearing 1-1 and 1-2 are selected as training set and the rest of them are used for testing. The full life diagram of raw signal is shown as Figure 5. The 500th-1000th collected data of bearing 1-1 and the 320th-400th collected data of bearing 1-2 are used as the normal stage data; the 2400th-2700th collected data of bearing 1-1 and the 831th-861th collected data are used as the data of severe fault stage.

Spectrum data is used as training data of binary classification neural network. Hardware of the experiment is a computer with i5-1035G1 CPU @ 1.00 GHz 1.19 GHz, 16 GB memory and software are MATLAB 2016a and PYTHON3.8.

After many attempts, the four-layer neural network is selected as the classifier, the number of the network nodes is 2048-10000-500-2 and the activation function of the front three-layer is the RELU function, the last layer use SoftMax function as activation function to implement the binary classification. The loss function is set as a cross-entropy function, train the network 20 times and the batch size is 8. In order to avoid false alarms, three consecutive predictions into the degradation stage mean that the stage is into degradation. Figure 6 shows some test bearing results. It can be seen from Figure 6 that the two-classification network can more accurately identify samples in the normal phase and samples in the degraded phase. Therefore, it can accurately determine the starting point of the degradation stage. The overall test results are shown in Table 2.

Table 2. Recognition of starting point during degradation phase

| Bearing | Failure time/s | Failure point/s |
|---------|----------------|-----------------|
| 1-1 | 2803 | 1517 |
| 1-2 | 871 | 821 |
| 1-3 | 2375 | 1332 |
| 1-4 | 1428 | 1090 |
| 1-5 | 2463 | 2444 |
| 1-6 | 2448 | 2100 |
| 1-7 | 2259 | 2241 |

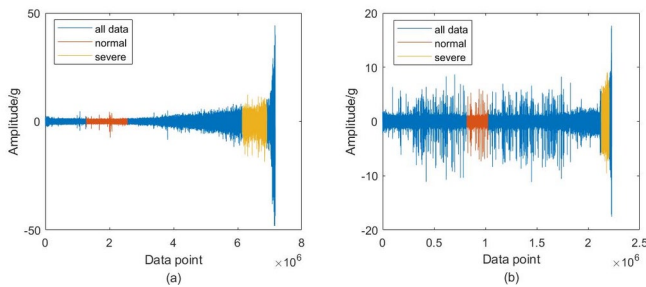


Fig. 5. The original vibration waveform of the bearing, (a) bearing 1-1, (b) bearing 1-2

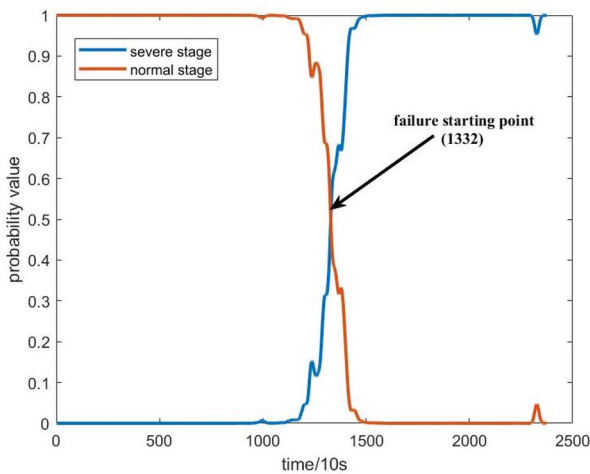


Fig. 6. Stage identification effect diagram of bearings 1-3

3.3. Prediction of RUL

3.3.1. Evaluation index and sample label

In order to quantitatively evaluate the effectiveness of the predictive RUL method proposed in this paper, this paper uses Root-Mean-Square-Error (RMSE) and Mean-Absolute-Error (MAE) as evaluation indicators. The calculation formula is shown in formula (11):

$$\begin{cases} MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i| \\ RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \end{cases} \quad (11)$$

where y_i is the actual value, \hat{y}_i is the predicted value, and m is the number of samples.

Since the prediction model of RUL used in this paper is a supervised learning model, it is necessary to label the source domain samples. This article uses the remaining life percentage of the bearing as the label for these samples. This tag can control the amount of data used for network training not to be too large, and improve computational efficiency. (For example, assuming failure time of bearing is 2500 s and time of degradation is 500 s, when the bearing running at 1500 s, the label for that point is $(\frac{1500-500}{2500-500}) = 50\%$).

3.3.2. Hyperparameters of the network

In order to obtain the best model prediction effect, this section discusses the important hyperparameters and network structure of the network. Since the setting of the learning rate will affect the convergence of the network model, which in turn affects the training effect of the model, the learning rate is an indicator that must be considered. Secondly, this paper uses the MMD function value as a scale function to measure the data of different failed bearings, and uses it as a part of the loss function, so it is of great significance to choose the MMD term trade-off coefficient. Therefore, this paper chooses the learning rate and the trade-off coefficient for experiments, and the selection range of hyperparameters is shown in Table 3.

Table 3. Value range of Hyper-parameters

| Hyperparameters | Range |
|-----------------|---|
| Learning rate | 0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001 |
| Trade-off value | 0, 0.1, 0.2, 0.3, 0.5, 0.7, 10, 50, 100 |

When the fixed trade-off coefficient is 0.2, try to experiment with different learning rate values. The prediction results are shown in Table 4. The values in Table 4 are the average values of multiple prediction results of all training set bearings. It can be seen from Table 4 that when the learning rate is large, the effect of the model is the worst. The possible reason is that a higher learning rate will prevent the network from converging to an optimal value. Because the gradient descent step is too large, it can only make the model hover around the optimal value, resulting in lower prediction accuracy. As the learning rate decreases, the prediction accuracy continues to improve. However, too small a learning rate will reduce the convergence speed. Under the same number of iterations, too small a learning rate may not achieve convergence. Therefore, considering the prediction accuracy and time-consuming considerations, this paper chooses the learning rate to be 0.001.

Table 5 shows the prediction effect of the compromise coefficient under different values. It can be seen from Table 5 that when the trade-off coefficient is selected as 0.2, the performance of the network model is the best. If the trade-off coefficient is too small, the constraint in-

Table 4. Influence of different learning rates on the prediction model

| Learning rate | MAE | RMSE |
|---------------|---------------|---------------|
| 0.1 | 0.2518 | 0.2909 |
| 0.01 | 0.1803 | 0.2219 |
| 0.005 | 0.1870 | 0.2313 |
| 0.001 | 0.1702 | 0.2085 |
| 0.0005 | 0.1930 | 0.2340 |
| 0.0001 | 0.1905 | 0.2280 |

Table 5. Influence of different trade-off coefficients on model prediction

| Trade-off value | MAE | RMSE |
|-----------------|---------------|---------------|
| 0 | 0.2121 | 0.2595 |
| 0.1 | 0.1923 | 0.2378 |
| 0.2 | 0.1702 | 0.2085 |
| 0.3 | 0.1858 | 0.2263 |
| 0.5 | 0.1876 | 0.2247 |
| 0.7 | 0.1992 | 0.2410 |
| 10 | 0.1999 | 0.2400 |
| 50 | 0.1909 | 0.2310 |
| 100 | 0.1999 | 0.2361 |

formation between different data sets will be reduced, and the model will not be able to learn domain-invariant features. When the trade-off coefficient is greater than 0.5, because the weight of the MMD term is too large, the loss of the prediction model cannot be trained well. In summary, the compromise factor of 0.2 in this article is reasonable.

In order to determine the influence of the architecture of the network model, this paper adds the MMD function to the last layer of the network model (MMD1), adds the MMD function to the penultimate layer (MMD2), and adds MMD function to the last two layers (MMD12). The experimental results are shown in Table 6. Since the network model extracts the shallow information of the network model in the first few layers, the features extracted by the network model are more abstract in the subsequent layers. It can be seen from Table 6 that the effect of the single-layer MMD function is not as good as that of the double-layer MMD function. This is mainly because the single-layer MMD function is not enough to represent the difference in data distribution between the training set and the test. Therefore, it is reasonable to choose MMD12 as the network model architecture of this article.

Table 6. Influence of different locations of MMD on prediction

| Trade-off value | MAE | RMSE |
|-----------------|---------------|---------------|
| MMD1 | 0.1723 | 0.2159 |
| MMD2 | 0.1860 | 0.2254 |
| MMD12 | 0.1702 | 0.2085 |

3.3.3. Prediction of RUL

The PHM data set is used as the analysis data to verify the effectiveness and feasibility of the method in this paper. The original data of bearing 1-1 is used as the training set, and the lifetime percentage is used as the sample label, which belongs to the source domain. Unmarked data for bearings 1-5 and 1-7 are used as auxiliary data. The test sets are Bearing 1-2, 1-3, 1-4, 1-6.

Through theoretical analysis and experimental verification, the hyperparameters of the experimental model are set that Optimizer is Adam, Learning rate=0.001, Trade-off=0.2, Epoch=400, Batch-

size=32. The network adopts two convolution and pooling layers for feature extraction, the kernel size is 5 in convolution and 2 in max pooling. In the transfer part of the full connection layer, the RBF function is selected as the kernel function for calculation of MMD distance and the width of the kernel is 1000. When the MMD measurement loss function accounts for 0.20 total loss, the network reaches the optimal effect. The batch size is 32, and half the data comes from the source domain, the rest is from the target domain. The epoch is set as 400. The loss function of the training process is shown in Figure 7. It can be seen that as the number of epochs increases, the loss of the training model does not decrease, indicating that the model has reached the effect of convergence. The prediction effect of the training set direction is shown in Figure 8. It can be seen from Figure 8 that this method shows a good fitting effect and good monotonicity for the bearings of the training set, and the failure time of the bearing can be almost perfectly predicted in the final stage. At the same time, it shows that the network architecture and hyperparameters selected in this paper are reasonable, and the network model can learn bearing degradation information from the training set.

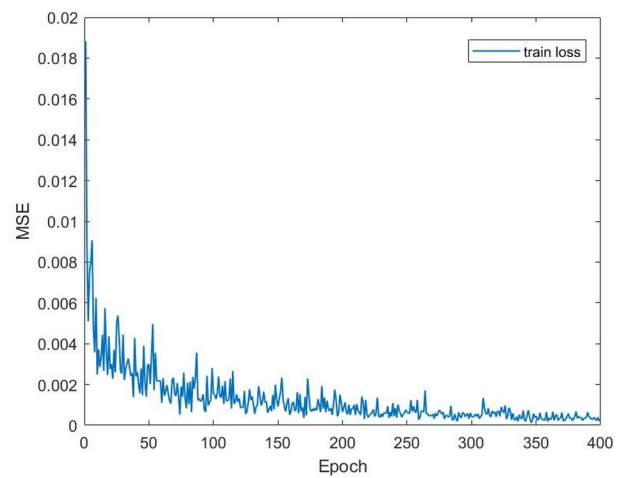


Fig. 7. Training loss diagram of network model

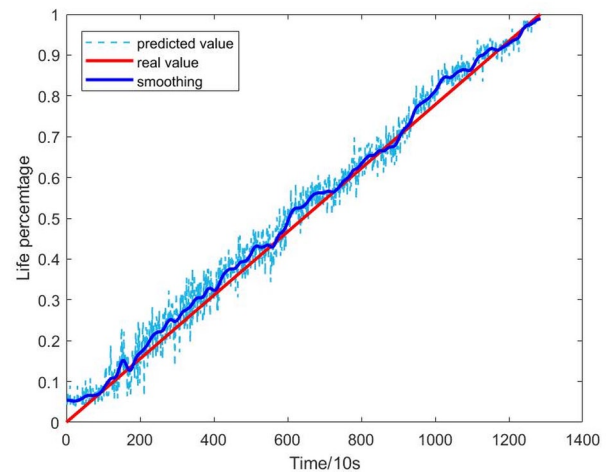


Fig. 8. The prediction effect of bearing in train set (bearing 1-1)

As shown in Figure 9, it can be seen that the method in this paper shows high prediction accuracy for both the suddenly failed bearing 1-2 and the gradually failed bearing 1-3, and the fluctuation of the predicted value is significantly reduced after sliding average processing. Although in the process of predicting the degradation trend of the network model, the monotonicity of the bearing 1-2 is not satisfactory. However, in actual engineering, people pay more attention to the degradation trend and final RUL value of the bearing in the later period of

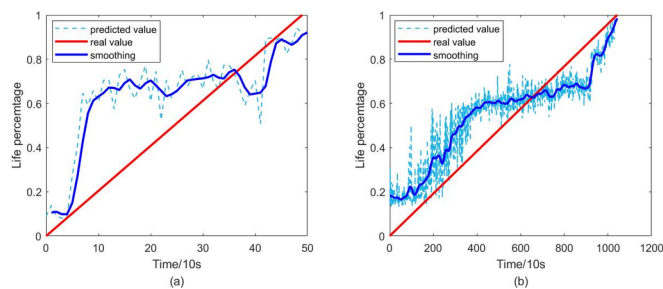


Fig. 9. The prediction effect of bearing in test set, (a) bearing 1-2; (b) bearing 1-3

operation. It can be seen from Figure 9 that both bearing 1-2 and bearing 1-3 have good monotonicity and higher prediction accuracy at the final moment. Even bearings 1-3 can predict the failure time almost without error at the last moment. In summary, the method proposed in this paper can meet the requirements of the RUL prediction of the bearing in actual engineering

3.4. Comparison analysis of model advantage

In order to verify the superiority of this method, this paper chooses the CNN model and the SPP-CNN model as the comparison model to verify the effectiveness of the improved strategy. Secondly, in order to verify the effectiveness of the migration strategy in this article, the current advanced migration learning models Transfer Component Analysis (TCA) and Domain-Adversarial Training of Neural Networks (DANN) are used as comparison models. The comparison model introduction is shown in Table 7.

Table 7. Comparison model

| Model | Input | Transfer method |
|---------------------------------|---------------------|----------------------|
| CNN | frequency spectrum | None |
| SPP-CNN | frequency spectrum | None |
| TCA [17] | traditional feature | MMD |
| DANN [7] | frequency spectrum | adversarial strategy |
| SPP-CNNLTL (Proposed method) | frequency spectrum | MMD |

In order to ensure the accuracy of the comparison effect, the architecture and hyperparameter settings of the comparison model are consistent with the selection of the proposed method. The experimental prediction results of different models are shown in Table 8.

Table 8. The MAE value of different models

| Model | bearing1-1 | bearing 1-2 | bearing 1-3 | bearing 1-4 | bearing 1-6 |
|------------|---------------|---------------|---------------|---------------|---------------|
| CNN | 0.0160 | 0.2828 | 0.2595 | 0.2083 | 0.3062 |
| SPP-CNN | 0.419 | 0.2580 | 0.1454 | 0.1651 | 0.3062 |
| TCA | 0.5023 | 0.2543 | 0.2034 | 0.1823 | 0.3124 |
| DANN | 0.0432 | 0.2392 | 0.1224 | 0.1523 | 0.3034 |
| SPP-CNNLTL | 0.0201 | 0.1802 | 0.1115 | 0.1332 | 0.2477 |

From Table 8, compared with other models, there are three kinds of advantages in the proposed method in this paper.

1. The SPP-CNN model improves the accuracy of bearing RUL prediction. Although the traditional CNN model has higher prediction accuracy on the training set, its prediction effect on the test set is worse than that of the SPP-CNN model. The main reason is that SPP can improve the generalization ability, thereby improving the RUL prediction effect of the bearing under different failure degradation.

2. Transfer learning improves the accuracy of bearing RUL prediction. After using transfer learning, the model prediction ability of the training set and test set has been improved. It also has a better predictive effect for bearings that suddenly fail.
3. In order to demonstrate the superiority of the transfer strategy, this paper chooses TCA and DANN as the comparison model. The TCA model maps the features of the source domain and the target domain to the high-dimensional replicable kernel Hilbert space to minimize the distance between the source domain and the target domain. The input of the TCA model is 24 traditional statistical features, including time-domain features and wavelet packet energy. It selects the RBF function as the kernel function. The DANN model uses domain confrontation strategies to solve the problem of data distribution differences. The prediction effect of each model is shown in Table 7. It can be seen from the evaluation indicators in Table 7 that this paper has a higher RUL prediction accuracy for the tested bearing. Compared with other transfer learning models, the proposed method has higher prediction accuracy. The main reason is the use of adaptive technology to solve the problem of inconsistent allocation between training data and test data. And use the SPP-CNN layer to improve the generalization ability of the network to obtain a better transmission effect.

4. Conclusion

This paper proposes a RUL prediction model of bearing based on multi-feature deep convolution transfer learning. First of all, this paper uses the SPP layer to avoid the problems of poor prediction accuracy and poor generalization ability of a single feature. Then, based on the MMD migration mechanism, the SPP-CNN model was improved, and the problem of inconsistent data distribution of the degradation trend of each bearing caused by the failure of each bearing was solved. Finally, by using the PHM2012 bearing public data set, and comparing the results with the prediction effect of the transfer learning model, the following conclusions are drawn: 1. The method proposed in this paper has good monotonicity in the final stage of various types of failed bearings. Higher prediction accuracy can meet the actual needs of engineering applications. 2. The domain adaptive module can reduce the data distribution difference between different failure trends, so that the model in this paper has a wider application range. From the above content, it can be seen that compared with the current advanced RUL prediction, the method in this paper has obvious advantages.

Considering the great potential of deep learning models in RUL prediction, future work shows that the RUL prediction of bearings under different working conditions should be considered, so that the RUL prediction model has stronger practicability.

Acknowledgement

This research is subsidized by the Natural Science Foundation of China, 'Research on reliability theory and method of total fatigue life for large complex mechanical structures' (Grant No. U1708255).

References

1. Ali J B, Chebel-Morello B, Saidi L, Malinowski S, Fnaiech F. Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network. *Mechanical Systems and Signal Processing* 2015; 56-57:150-172, <https://doi.org/10.1016/j.ymssp.2014.10.014>.
2. Burns J E, Yao J, Chalhoub D, Chen J J, Summers R M. A Machine Learning Algorithm to Estimate Sarcopenia on Abdominal CT. *Original Investigation* 2020; 27(3):311-320, <https://doi.org/10.1016/j.acra.2019.03.011>.
3. Cheng H, Kong X, Chen G, Wang Q, Wang R. Transferable convolutional neural network based remaining useful life prediction of bearing under multiple failure behaviors. *Measurement* 2020; 168:108286, <https://doi.org/10.1016/j.measurement.2020.108286>.
4. Costa P, Akcay A, Zhang Y, Kaymak U. Remaining Useful Lifetime prediction via deep domain adaptation. *Reliability Engineering & System Safety* 2020; 195:106682, <https://doi.org/10.1016/j.res.2019.106682>.
5. Dong S, Luo T. Bearing degradation process prediction based on the PCA and optimized LS-SVM model. *Measurement* 2013; 46(9):3143–3152, <https://doi.org/10.1016/j.measurement.2013.06.038>.
6. Dong S, Wen G, Lei Z, Zhang Z. Transfer learning for bearing performance degradation assessment based on deep hierarchical features. *ISA Transactions* 2020; 108(9):343-355, <https://doi.org/10.1016/j.isatra.2020.09.004>.
7. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V. Domain-adversarial training of neural networks. *The journal of machine learning research* 2016; 17(1): 2096-2030, <https://dl.acm.org/doi/abs/10.5555/2946645.2946704>.
8. Guo L, Lei Y, Xing S, Yan T, Li N. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics* 2018; 66(9): 7316-7325, <https://doi.org/10.1109/TIE.2018.2877090>.
9. Hinch A Z, Tkouat M. Rolling element bearing remaining useful life estimation based on a convolutional long-short-term memory network. *Procedia Computer Science* 2018; 127:123-132, <https://doi.org/10.1016/j.procs.2018.01.106>.
10. LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1989; 1(4): 541–551, <https://doi.org/10.1162/neco.1989.1.4.541>.
11. Lei Y, Li N, Gontarz S, Jing L, Radkowski S, Dybala J. A model-based method for remaining useful life prediction of machinery. *IEEE Transactions on reliability* 2016; 65(3): 1314-1326, <https://doi.org/10.1109/TR.2016.2570568>.
12. Lei Y, Li N, Guo L, Li N, Yan T, Jing L. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing* 2018; 104:799-834, <https://doi.org/10.1016/j.ymssp.2017.11.016>.
13. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Matti P. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision* 2020; 128: 261–318, <https://doi.org/10.1007/s11263-019-01247-4>.
14. Lu W, Liang B, Cheng Y, Meng D, Yang J, Zhang T. Deep model based domain adaptation for fault diagnosis. *IEEE Transactions on Industrial Electronics* 2016; 64(3):2296-2305, <https://doi.org/10.1109/TIE.2016.2627020>.
15. Mao W, He J, Ming J Z. Predicting Remaining Useful Life of Rolling Bearing Based on Deep Feature Representation and Transfer Learning. *IEEE Transactions on Instrumentation and Measurement* 2019; 69(4):1594-1608, <https://doi.org/10.1109/TIM.2019.2917735>.
16. Nectoux P, Gouriveau R, Medjaher K, Ramasso E, Varnier C. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In: *IEEE International Conference on Prognostics and Health Management*. Denver, CO, USA, 1-8, 2012, <https://hal.archives-ouvertes.fr/hal-00719503>.
17. Pan S J, Tsang I W, Kwok J T, Yang Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 2011; 22(2):199–210, <https://doi.org/10.1109/TNN.2010.2091281>.
18. Mao W, He J, Ming J Z. Predicting Remaining Useful Life of Rolling Bearing Based on Deep Feature Representation and Transfer Learning. *IEEE Transactions on Instrumentation and Measurement* 2019; 69(4):1594-1608, <https://doi.org/10.1109/TIM.2019.2917735>.
19. Rai A, Upadhyay S H. A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings. *Tribology International* 2016; 96:289-306, <https://doi.org/10.1016/j.triboint.2015.12.037>.
20. Rai A, Upadhyay S H. Intelligent bearing performance degradation assessment and remaining useful life prediction based on self-organising map and support vector regression. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 2018; 232(6):1118-1132, <https://doi.org/10.1177/0954406217700180>.
21. Salakhutdinov R, Hinton G. An Efficient Learning Procedure for Deep Boltzmann Machines. *Neural Computation* 2012; 24(8): 1967–2006, https://doi.org/10.1162/NECO_a_00311.
22. Su C, Li L, Wen Z. Remaining useful life prediction via a variational autoencoder and a time-window-based sequence neural network. *Quality and Reliability Engineering International* 2020; 36(5): 1639-1656, <https://doi.org/10.1002/qre.2651>.
23. Tang Y, Chen M, Wang C, Luo L, Zou X. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. *Frontiers in Plant Science* 2020; 11:510, <https://doi.org/10.3389/fpls.2020.00510>.
24. Wang B, Lei Y, Li N, Li N. A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings. *IEEE Transactions on Reliability* 2020; 69(1):401-412, <https://doi.org/10.1109/TR.2018.2882682>.
25. Wang B, Lei Y, Yan T, Li N, Guo L. Recurrent convolutional neural network: A new framework for remaining useful prediction of machinery. *Neurocomputing* 2020; 379:117-129, <https://doi.org/10.1016/j.neucom.2019.10.064>.
26. Wang F K, Mamo T. Hybrid approach for remaining useful life prediction of ball bearings. *Quality and Reliability Engineering International* 2019; 35(7): 2494-2505, <https://doi.org/10.1002/qre.2538>.
27. Wang Q, Zhao B, Ma H, Chang J, Mao G. A method for rapidly evaluating reliability and predicting remaining useful life using two-dimensional convolutional neural network with signal conversion. *Journal of Mechanical Science and Technology* 2019; 33:2561–2571, <https://doi.org/10.1007/s12206-019-0504-x>.
28. Wang Y, Peng Y, Zi Y, Jin X, Tsui K L. A Two-Stage Data-Driven-Based Prognostic Approach for Bearing Degradation Problem. *IEEE Transactions on Industrial Informatics* 2016; 12(3): 924-932, <https://doi.org/10.1109/TII.2016.2535368>.
29. Ye Z S, Xie M. Stochastic modelling and analysis of degradation for highly reliable products. *Applied Stochastic Models in Business and Industry* 2015; 31(1):16-32, <https://doi.org/10.1002/asmb.2063>.
30. Zhang J, Wang P, Yan R, Gao R X. Long short-term memory for machine remaining life prediction. *Journal of Manufacturing Systems* 2018; 48(C): 78-86, <https://doi.org/10.1016/j.jmsy.2018.05.011>.
31. Zhang Y, Yang S, Li P, Hu X, Wang H. Marginalized Stacked Denoising Autoencoder with Adaptive Noise Probability for Cross Domain

- Classification. IEEE Access 2019; 7:2169-3536, <https://doi.org/10.1109/ACCESS.2019.2925811>.
32. Zhu J, Chen N, Peng W. Estimation of Bearing Remaining Useful Life based on Multiscale Convolutional Neural Network. IEEE Transactions on Industrial Electronics 2019; 66(4):3208-3216, <https://doi.org/10.1109/TIE.2018.2844856>.
33. Zhu J, Chen N, Shen C. A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions. Mechanical Systems and Signal Processing 2020; 139: 106602, <https://doi.org/10.1016/j.ymssp.2019.106602>.

Application of machine learning and rough set theory in lean maintenance decision support system development

Indexed by:



Katarzyna Antosz^{a,*}, Małgorzata Jasiulewicz-Kaczmarek^b, Łukasz Paśko^a, Chao Zhang^c, Shaoping Wang^c

^aRzeszów University of Technology, Faculty of Mechanical Engineering and Aeronautics, Powstańców Warszawy 8, 35-959 Rzeszów, Poland

^bPoznań University of Technology, Faculty of Management Engineering, Prof. Rychlewskiego 2, 60-965 Poznań, Poland

^cBeihang University (BUAA), School of Automation Science and Electrical Engineering, 37 Xueyuan Road, Beijing, 100191, China


Highlights

- A review of lean maintenance importance in manufacturing.
- A approach with rough set theory and decision tree.
- Rough set theory with different types of algorithms selected for predictive models.
- The classification model for lean maintenance implementation assessment.

Abstract

Lean maintenance concept is crucial to increase the reliability and availability of maintenance equipment in the manufacturing companies. Due the elimination of losses in maintenance processes this concept reduce the number of unplanned downtime and unexpected failures, simultaneously influence a company's operational and economic performance. Despite the widespread use of lean maintenance, there is no structured approach to support the choice of methods and tools used for the maintenance function improvement. Therefore, in this paper by using machine learning methods and rough set theory a new approach was proposed. This approach supports the decision makers in the selection of methods and tools for the effective implementation of Lean Maintenance.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

lean maintenance, availability, machine learning, decision trees, rough set theory

1. Introduction

Maintenance processes have a significant impact on manufacturing companies such as: production efficiency, safety and environment requirements and customers satisfaction [19, 23, 28, 38]. In addition, delivering high-quality products with tighter tolerances and lower waste and rework levels also depends on well-maintained equipment, which is another reason to develop more efficient maintenance processes [39]. Moreover, Marksberry [52] determined as the waste of production process the 'maintenance of machines and devices'. Various concepts have been used to decrease reliability and availability of machines and devices, one of them is Lean Maintenance (LMn) [29]. LMn deals with the integration of people in the production process, using certain methods and tools for continuous improvement, as well as the elimination of waste in value-added activities.

The complexity of various LMn tools and methods as well as the investment costs make the LMn implementation a difficult and complex process, although this concept has an impact on the business results of the organization [10]. The problem of inadequate understanding of the relationship between LMn and the operating environment of manufacturing companies causes the LMn implementation to fail [17]. Therefore, an important aspect is the development of systems supporting the assessment of the effectiveness of LMn implementation. [91].

The aim of the article is to develop a decision support system, which will be helpful for decision-makers from companies in select-

ing appropriate LMn methods and tools that have the greatest impact on the company's operational results. In the proposed decision making system the machine learning methods and rough set theory was used. The main research question was: Which of the LMn tools had the greatest impact on reducing the number of unplanned downtime?

The remainder of this paper is structured as follows. In Section 2 the literature review according the importance of maintenance function in manufacturing and lean maintenance is presented. Then, in Section 3 the research methodology is presented. In Section 4 the results of using decision trees and rough set theory to generate categorization models in the assessment of the implementation of lean maintenance are presented. Finally, the conclusions and direction of the future research are presented.

2. Background

2.1. The importance of maintenance function in manufacturing

Modern manufacturing companies focus on the availability, reliability and productivity of their manufacturing machines and devices [39, 84]. Equipment maintenance and system reliability are important factors that have impact on the ability to provide quality and timely products to clients, comply with legal requirements, and meet business goals. These needs have placed the maintenance function in the

(*) Corresponding author.

E-mail addresses: K. Antosz - kcktmip@prz.edu.pl, M. Jasiulewicz-Kaczmarek - malgorzata.jasiulewicz-kaczmarek@put.poznan.pl, Ł. Paśko - lpasko@prz.edu.pl, C. Zhang - cz@buaa.edu.cn, S. Wang - shaopingwang@vip.sina.com

spotlight as a strategic function for manufacturing companies [54, 58, 78, 79].

As defined by European standard EN 13306, maintenance is a “the combination of all technical, administration and management actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can deliver the required function (function or a combination of functions of an item which are considered necessary to provide a given service).” The presented definitions express the multidisciplinary character of maintenance operations, which include both technical aspects of the technical facility performance and all in-service aspects, referring to the facility itself and to all stakeholders and resources engaged into maintenance processes. According to [66] “Maintenance operations are much like manufacturing operations where both employ processes that add value to the basic input used to create the end product”

As maintenance management in a manufacturing company combines various functions (organizational and business) its implementation is complex and requires the utmost attention. According to [89] “maintenance is not just ensuring healthiness of equipment in a facility but it also plays a crucial role in achieving organization’s goals and objectives with optimum maintenance cost and maximum production. [...] and needs to be viewed as a strategic function in an organization”. Defining an appropriate maintenance strategy is seen as a way to turn your company’s goals into maintenance goals [89]. Maintenance objectives at strategic and tactical levels of the organization can be define in five categories [88]. First category is maintenance budget, which consists e.g. maintenance costs and maintenance value. In the second category functional and technical aspects such as: availability, maintainability, reliability, Overall Equipment Effectiveness (OEE), productivity, maintenance and output quality are described. Third category contains plant design life. Next category includes inventory of spare parts and logistics. Finally, people and environment are counted in the last category. To achieve this objectives maintenance strategies have evolved with the course of time, From reactive maintenance (“run-to fail” logic) to proactive maintenance (PrM) strategies such as: Preventive Maintenance (PM) or Predictive Maintenance (PD). The main goal of PrM strategies is to monitor the equipment and making minor repairs to keep them in the good condition with high performance. Research conducted by [32] shows that adopting predictive maintenance in an enterprise can minimize maintenance costs up to 30% and eliminate breakdowns up to 75% compared to preventive maintenance.

Today, maintenance with a strategic role in revenue generation is seen as source of added-value, with key role for driving performance improvement [51]. According to [37] “advanced practice in maintenance can play a role in achieving more competitive, responsible and sustainable performance in manufacturing companies.” In this line maintenance should be view as an important function in achieving sustainability in manufacturing processes. Many researcher start to study the impacts and contributions of maintenance function to more sustainable operations in manufacturing companies. From the economic dimension of sustainable manufacturing four factors quality and productivity, delivery on time, innovation and cost are affected by the maintenance function [40, 49]. From environment dimension of sustainable manufacturing most frequently prevention of environmental damage, emissions reduction and land conservation, energy consumption reduction and energy savings are underlined [60, 72]. Finally, from the social dimension of suitability manufacturing are underlined the relationship of the maintenance function with its stakeholders within and outside the company, with a particular focus on the maintenance personnel, who is affected by decisions made in the maintenance department [22, 41].

Manufacturing industry has now embarked on a digital transformation following the Industry 4.0 paradigm in which the maintenance organization is expected to play a key role in enabling robust autonomous systems [49]. According to [56], many companies consider maintenance processes improvement as the one of the initial stages towards Industry 4.0 concept.

The growing complexity of the production environment, new requirements and new opportunities force the maintenance managers to constantly search for opportunities to improve activities and processes. Dekker [27] stress that “the main question faced by maintenance management, whether maintenance output is produced effectively, in terms of contribution to company profits”. Although this question was asked many years ago, it is still timely and is very difficult to answer. Many researchers and practitioners proposed models to solve maintenance-related problems and pointed out that successful implementation of these models depends on appropriate understanding and using properly tools and techniques indicated in this models.

2.2. Lean and maintenance

Lean Manufacturing (LM) is worldwide recognition methodology for the improvement of internal processes, popularised by the book ‘The Machine that Changed the World’ [15]. The main challenge of LM is to increasing customer satisfaction while decreasing waste and losses. The benefits of lean implementation are divided in two field. Firstly, LM eliminates wastes, decreases delivery, lead and cycle times, decrease inventories, and increase the productivity [11, 45]. Secondly, LM improves the workers satisfaction, good communication, and decision-making process [25].

LM demand for a reliable and stable machine operation gave way to another concept - Lean Maintenance [82] also known as Lean TPM (Total Productive Maintenance) [55]. According to [82] “without a Lean Maintenance operation, Lean Manufacturing can never achieve the best possible attributes of Lean”, so “first – Lean Maintenance, and next – Lean Manufacturing”.

According to [77] “Lean production shifts the attention of maintenance improvement from the technical matters to the management side, which focuses on eliminating the root causes of problems through team-based decisions and implementation”.

Smith and Hawkins [82] defined LMn as “proactive maintenance operation employing planned and scheduled maintenance activities through total productive maintenance (TPM) practices, using maintenance strategies developed through application of reliability centered maintenance (RCM) decision logic and practiced by empowered (self-directed) action teams using the 5S process, weekly Kaizen improvement events, and autonomous maintenance together with multi-skilled, maintenance technician-performed maintenance through the committed use of their work order system and their computer maintenance management system (CMMS) or enterprise asset management (EAM) system”. This definition extends beyond the classic LM concept of TPM including a reliability approach based on the RCM method. It indicates the need to identify hazards, assess their consequences and on this basis, determine the criticality of technical facilities and appropriate maintenance activities for the function performed by the facility.

LMn is based on a multidimensional management concept focused on the waste and losses elimination. [26]. Each maintenance operation is associated with unwanted side effects and wastes, such as [35]: (1) Over-maintenance; (2) Waiting for resources; (3) Task sequencing and scheduling; (4) Maintenance task processing; (5) Excessive inventory; (7) Motion; (8) Correction.

One of the main steps for improving the maintenance processes is to develop a system to identify VA (Value Added) and NVA (Non Value Added) activities and recognize the types wastes [76]. To achieve this LMn includes several tools and methods, such as: 5S, Value Stream Mapping (VSM), Single Minute Exchange od Die (SMED), TPM, Visual Management (VM) (Figure 1).

These methods and tools simplified maintenance processes and improve the maintenance performance.. The reduction of waste in maintenance means a reduce setup time and increase OEE [9, 57, 92], better management of consumable materials and spare parts [68], downtime reduction [36, 85] and lower the Mean Time To Repair (MTTR) and standardization of maintenance procedures [29]. Barnard [12] pointed out that lean can help to develop Reliability Pro-

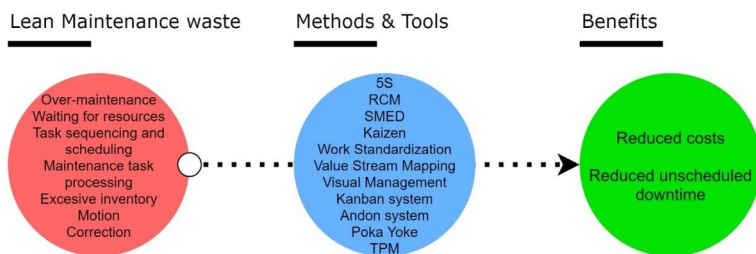


Fig. 1. From waste to benefits – Lean Maintenance perspective

gram Plan and to select only VA activities for execution. In the work [53] the authors suggest how LM principles can be adopted to LMn and underlined the importance of data in maintenance management process decision-making.

Evidence of LMn tools implementation is found in various sectors, such as the automotive industry [6, 67], aerospace industry [21], power plants [29] textile industry [7, 73], food industry [8], oil and gas industry [24, 76] among others. Such evidence points to a number of universality and the use of the LMn tools in different contexts and companies, increasing its importance as an approach to continuous improvement [30, 70, 86]. However, implementation of LMn tools / practices is time consuming and costly process and needs continuous efforts to get effective results. Furthermore, there is no roadmap, no unified model and standard answer on the way to achieve lean [93].

Many researchers are identified industrial problems regarding LM implementation [1, 59]. To support practitioners in effective implementation of LM methods and tools, various models suitable for different industries were developed [2, 16]. For selection of lean tools in a manufacturing organisation [47] propose fuzzy FMEA, AHP and QFD-based approach, [61] use of AHP method and illustrate based on example related to the construction works, [42] proposes the improved VIKOR method and idea of multiple criteria decision-making for LM tool selection, [80] applies grey method for LM tool selection.

The above analyzes show that the choice of LM practices is not a simple problem. Moreover, the benefits of implementing LM may be different [86]. Since maintenance management in manufacturing companies connects various function (organizational and business) and activities, LMn methods tools implementation is complex and requires knowledge and skills. Maintenance managers, a specially in small in medium-sized enterprises, have a problem of selecting the best in a given operational context of the enterprise. Thus, development of decision-making support tools can assist in LMn tools performance appraisal, facilitating appropriate LMn practices [29].

3. Research methodology

The purpose of this research was to identify the main factors impacting on effectiveness of LMn implementation in manufacturing companies. To archive this goal the machine learning (ML) method and rough set theory (RST) was proposed.

The research methodology consist of two stages. The first stage presents the results of the study, conducted in the manufacturing companies, concerning the maintenance management and lean tools implementation. Then, the obtained data was pre-proceed and statistical analyses was performed (Section 3.1)

In the second stage firstly the data set was divide into two sets: training and test data set. Then the decision trees (DT) (Section 3.2) and RST (Section 3.3) to generate the decision rules were used. The main goal of this stage was to generate the decision rules, which shows the relationships between the activities undertaken as part of the implementation of the lean maintenance concept and the results achieved. DT and RST were used for the variable of the number of unplanned downtime (NUD) indicator. Finally the obtained results were compared (Section 3.4). The detailed research methodology on Figure 2 is presented.

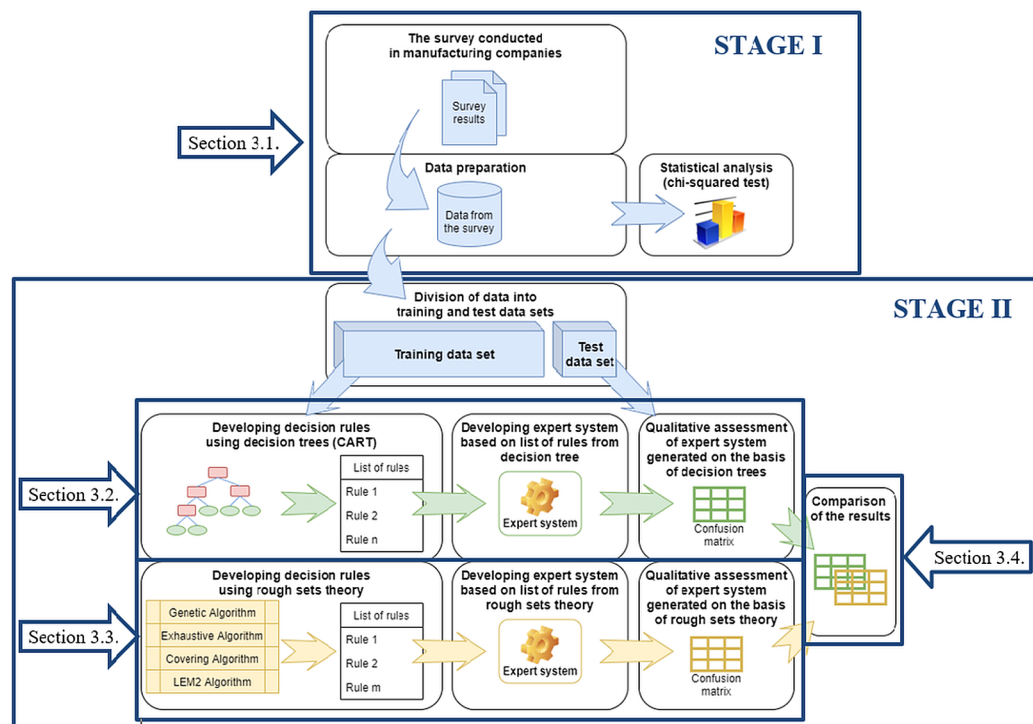


Fig. 2. The research methodology

3.1. Data collection and preliminary analysis

In the first stage the data for the research in manufacturing companies were collected. For participation in this research the companies of various sizes and from various industries were invited. The research involved companies that had been implementing the LMn concept for at least 5 years such as SMED, TPM, 5S. For the research the survey method was used. The research involved mainly representatives of top and middle management as well as employees directly related to the supervision of the maintenance process in the company. An important element of the research was to obtain information about the types of benefits identified by enterprises after the implementation of LMn tools such as: TPM, 5S and SMED. The obtained data from the survey was adequately prepared. The first stage was their pre-processing, which included data selection and cleaning. The purpose of this step was to remove inconsistent or erroneous data. In the data preparation the processing technique by removing the missing data was used. This had the effect of reducing the size of the dataset After then, the statistical analyse for identification the factors which have the impact on the NUD value in surveyed companies was used. In Section 4.1 and 4.2 the results of the first stage of the research are presented.

3.2. Machine learning and decision trees

In the second stage firstly the pre-proceed data set was divided into two sets: training and test data set. The training data set for developing the classification models was used. However, the test set for them validation was used. Firstly the machine learning method (decision trees) for generating the decision rules (classification model) was used.

ML combines solutions from the fields of statistics, computer science, cognitive sciences, recognition theory and many other fields [14]. Developed in the nineties of the last century, data mining methods are one of the most widely used IT tools at the present time [33]. These methods are included in modern applications. Moreover, these methods are used by the middle and top management level to make decisions based on the knowledge “retrieved” from the internal documentation of the organization and the results of the conducted research. The use of machine learning methods is divides in three stages: data preparation, data analysis (model building) and implementation. ML methods were successfully implemented in many different areas [14, 65] also in maintenance management [43, 46, 74, 87, 90].

One of the ML methods used for constructing the models are DT. DT are the one of the most popular and effective methods of ML [13]. DT are built mostly recursively (top-down approach) [34, 71].

DT construction is performing by in-depth search of all available variables and all possible splits in the data set for each decision node (t) by choosing the optimal partition [48]. $\{(y_i, x_i)\}_{1 \leq i \leq n}$ denotes the analysed data set, where $y_i \in \{c_1, c_2, \dots, c_s\}$ and $x_i = (x_{i1}, x_{i2}, \dots, x_{ik}) \in R^k$. The values c_1, c_2, \dots, c_s means possible classes characteristic y . The task of classification consists is to divide space R^k on q separated areas, where each area corresponds to a certain class. Based on the ob-

servation of the characteristics $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ can be analyzed object classification [3].

In this study the Classification and Regression Trees (CART) algorithm was used. This algorithm is one of the basic algorithms proposed by [18]. The Gini index, also called as the impurity measure, has been proposed by the authors of the algorithm. The entire space R^k is divided into q separated regions, $R_1 \cup R_2 \cup \dots \cup R_q = R^k$. For the node m , $1 \leq m \leq q$, representing region R_m , the Gini index is determined as follows (1) [3]:

$$Q_G(m) = \sum_{j=1}^s p_{mj}(1 - p_{mj}) = 1 - \sum_{j=1}^s p_{mj}^2 \quad (1)$$

where p_{mj} is a conditional probability for j -th class in a node, s – a number of classes. In node m with n_m observations the conditional probability for j -th class is equal (2):

$$p_{mj} = \frac{\#\{y = c_j : x \in R_m\}}{n_m} \quad (2)$$

The decision rule generated by CART algorithm were used to develop an expert system (with the use of PC-Shell /Aitech Sphinx). In the system, for creation the knowledge base two blocks: facet and rules were used. For declaration the values and attributes of decision the facets block was used. In the decision nodes the explanatory variables as decision attributes were placed. The target attribute represented the results of system inference. The NUD value was finally obtained in a separate output window.

For validation the developed decision rules in the expert system the data from companies (test set data) was used. Then, the confusion matrix and k-fold cross-validation to assess the quality of developed DT was used. In the confusion matrix the following values were determined: TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). To assess the quality of the developed classifier the indicators proposed by [31, 62, 83] were used (Table 1).

In Section 4.3 the results of the this step of the research are presented.

3.3. Rough Set Theory

For developing the second classification model the RST was used. This theory is recognized as a tool that allows to reduce the input dimension and finds a way to reduce the uncertainty and ambiguity of data. Recently, there has been a very rapid development in this area and the possibilities and application of this theory in ML and decision-making systems. [50, 64, 81]. The main advantage of this theory is the ability to find the relationship between the explanatory variables and the dependent variables, which allows to support the decision-making process based on data analysis. Moreover, RST allows for dimensionality reduction (elimination of explanatory variables that have no influence on the explained variables). Knowledge extracted using RST is generated in the form of decision rules [50].

The formal description of the rough set theory in the works [63, 64] is presented. In order to start data analysis using this theory, the concept of an information system and a decision table should be defined. Let S be a decision system define as $S = \langle U, A, V, f \rangle$. Where U is a non-empty, finite

Table 1. Indicators – quality of DT

| No. | Indicator | Formula |
|-----|-----------------------------------|---|
| 1. | Accuracy (Acc) | $Acc = \frac{TP + TN}{TP + TN + FP + FN}$ |
| 2. | Overall error rate (Err) | $Err = \frac{FP + FN}{TP + TN + FP + FN}$ |
| 3. | True positives rate (TPR) | $TPR = \frac{TP}{TP + FN}$ |
| 4. | True negatives rate (TNR) | $TNR = \frac{TN}{TN + FP}$ |
| 5. | Positive predictive value (PPV) | $PPV = \frac{TP}{TP + FP}$ |
| 6. | Negative predictive value (NPV) | $NPV = \frac{TN}{TN + FN}$ |
| 7. | False positive rate (FPR) | $FPR = \frac{FP}{FP + TN} = 1 - TNR$ |
| 8. | False discovery rate (FDR) | $FDR = \frac{FP}{FP + TP}$ |
| 9. | False negatives rate (FNR) | $FNR = \frac{FN}{TP + FN} = 1 - TPR$ |
| 10. | Matthew's corr. coefficient (MCC) | $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(FN + TN)(FP + TN)}}$ |
| 11. | F1-score (F1) | $F1 = \frac{2 \times PPV \times TPR}{PPV + TPR}$ |
| 12. | Youden's J statistic (J) | $J = TPR + TNR - 1$ |

set of n objects $\{x_1, x_2, \dots, x_n\}$, called the universe. A is a non-empty, finite set of m attributes $\{a_1, a_2, \dots, a_m\}$ which characterize the analyzed objects. On the other hand, $V = \bigcup_{a \in A} V_a$, where V_a is called the domain of the attribute $a \in A$, which contains the values of this attribute. In turn $f: U \times A \rightarrow V$ is an information function such that $\wedge_{a \in A, x \in U} f(a, x) \in V_a$. An information system (IS) is called a decision table DT when there are separate sets of conditional C attributes and decision attributes D such as: $C \cup D = A$ and $C \cap D = \emptyset$. Then the decision table DT is described as follows: $DT = \langle U, C, D, V, f \rangle$. Using the properties of RST allows for extending the possibilities of such a table, which leads to a significant simplification of the rules. Consequently, the decision-making system takes on the features of generalization and constitutes an effective and intelligent data processing tool. RST proposes to replace an imprecise concept with a pair of precise concepts, called the lower and upper approximation of this concept [69]. The difference between the upper and lower approximations is precisely the boundary area to which all cases belong that cannot be correctly classified on the basis of current knowledge. If $IS = \langle U, A, V, f \rangle$ is an IS such that $B \subset A$ and $X \subset U$ are: B^* – the lower approximation of the set X in the IS, is the set: $XB^* = \{x \in U : B(x) \subseteq X\}$; B^* – the upper approximation of the set X in the IS is the set: $XB^* = \{x \in U : B(x) \cap X \neq \emptyset\}$; B – positive area of the set X in the IS we call the set: $POS_B(X) = XB^*$; B – the boundary of the set X in the IS we call the set: $BN_B(X) = XB^* - XB^*$; B – a negative region of X in the IS is the set: $NEG_B X = U - XB^*$. The definitions formulate the following conclusions: $XB^* \subset X \subset XB^*$; X is B when: $XB^* = XB^* \Leftrightarrow BN_B X = \emptyset$ and X is B -approximate when: $XB^* = XB^* \Leftrightarrow BN_B X \neq \emptyset$.

The lower approximation of the concept is therefore the area that defines all the objects that there is no doubt that they represent the concept in the light of the possessed knowledge. The upper approximation includes objects that cannot be ruled out that they represent this concept [20]. The edges are all those objects for which it is not known whether or not they represent a given set. There is also the so-called a numerical characteristic of the approximation of a set, which, using the coefficient of accuracy of the approximation (approximation), allows us to quantitatively characterize the blurriness of concepts [44].

In this study the RST allowed to generate a set of decision rules that can be used to construct decision systems. They are usually created in four iterative steps: identification of possible sets of values, isolation of conditional attributes (premises) and decision attributes, creation of decision rules in the form of IF - THEN, implementation in the decision system.

As in the case of DT the developed decision rules were implemented in the expert system. Moreover, the data test set to validate the decision rules and to assess the quality of the classifier the same indicators were used. The results of this step of the research in Section 4.4 are presented.

3.4. Comparison of the results

In the last step of the research the comparison of the results obtained from the assessment of developed classification models by DT and RST was performed. In the comparison the value of the indicators for DT and RST (Table 1) was analyzed. The analyses for the most frequently occurring classes was performed. In Section 4.5 the results of this step of the research are presented.

4. Results and analyses

4.1. The structure of the surveyed companies

The research was carried out in manufacturing companies in Podkarpackie Voivodship (Poland). The companies participating in the

study used various methods and tools of LMn. Figure 3 shows the percentage of surveyed companies implemented various tools of LMn.

The research was carried out in manufacturing companies in Podkarpackie Voivodship (Poland). The companies participating in the study used various methods and tools of LMn. Figure 3 shows the percentage of surveyed companies implemented various tools of LMn.

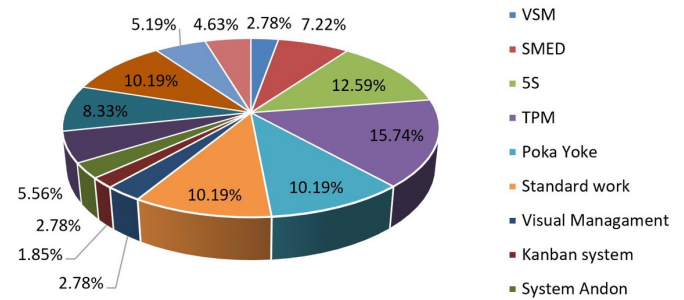


Fig. 3. Structure of the companies - LMn implementation

The surveyed companies were classified, inter alia, according to the following criteria: size of the organization, type of production, type of industry, and maintenance strategy. In the research the biggest group were large companies (70.77%) and companies from aviation industry (41.54%) and also companies with large batch production (25.68%) (Fig. 4, 5 and 6).

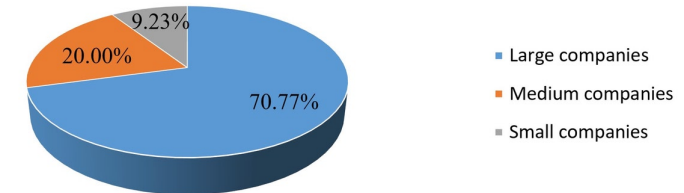


Fig. 4. Structure of the companies - the size of the company

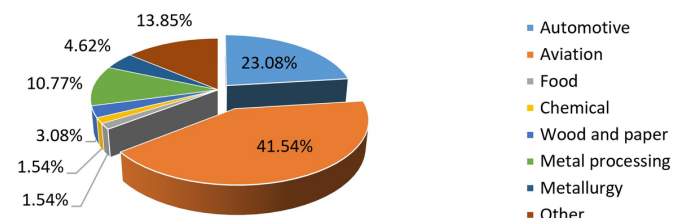


Fig. 5. Structure of the companies - the type of industry

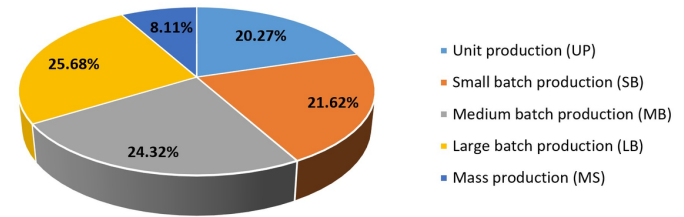


Fig. 6. Structure of the companies - type of production

In the analysed companies dominated preventive maintenance (PM) strategy, in particular: maintenance scheduled inspections (PM), maintenance scheduled inspections and repairs (PM) and autonomous maintenance (AM) (Fig. 7).

The implementation of the TPM system in the production plant significantly facilitates the process of supervising machines and technological devices. The main benefit of implementing TPM is the awareness of employees who, in conflicts and accompanying problems, find opportunities for continuous improvement. The decisive role in

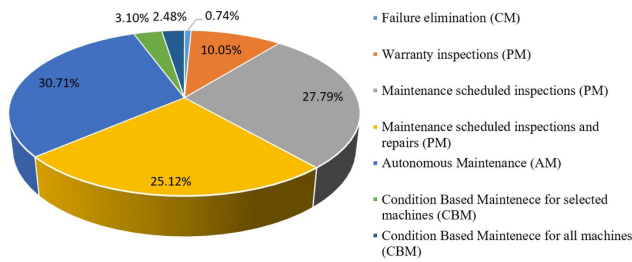


Fig. 7. Structure of the companies - maintenance strategy

assessing the effectiveness of TPM implementation in the enterprise allows the ongoing monitoring of the effects of TPM implementation. Many of the surveyed companies emphasized that the main effect is to reduce the number of unplanned downtime (UD). Any sudden shutdown of a machine from the production process was called an unplanned downtime. The most common reason for such a downtime is a mechanical, electrical or electronic failure, which poses a risk to safety at the workplace and failure to maintain proper operating parameters. To assess the effectiveness of the implementation of the LMn concept, enterprises used mainly OEE indicator and the number of unplanned downtimes (NUD). The research results concerning OEE are presented in the work [4]. This paper presents the results of the impact of LMn concept implementation on reducing the number of unplanned downtimes (NUD).

Figures 8 and 9 show the effects of implementing the LMn system – decreasing of NUD, in the surveyed companies. The analysis of this indicators was based on the following criteria: enterprise size and industry. When analyzing the results presented in Fig. 6, it should be noted that in the surveyed companies, the implementation of LMn most often resulted in a reduction of NUD in the range of 10-30% in the case of medium and large companies. The least, however, is above 50%. Small companies most often reported a reduction of NUD of less than 10%.

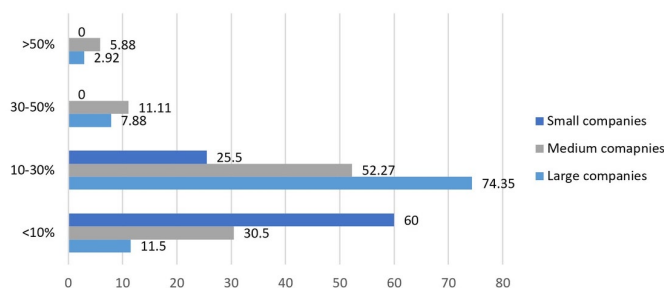


Fig. 8. The effects of implementing the LMn system (decreasing of NUD) – size of the company

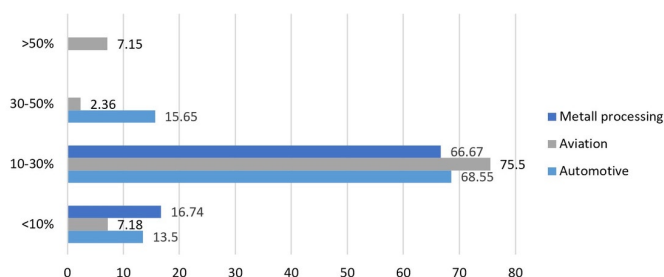


Fig. 9. The effects of implementing the LMn system (decreasing of NUD) – industry

The companies from various industries most often indicated a reduction in unplanned downtime also in the range of 10-30%. In 7.15%

of the aviation industry enterprises, NUD indicator is reduced by more than 50%.

4.2. Statistical analyses

Table 2 presents the analyzed factors which have potential influence on NUD indicator and the p-value.

Table 2. Potential factor influencing on NUD indicator and p-value

| Number | Factor | p-Value |
|--------|--|---------|
| | | NUD |
| 1 | The size of the company | 0.318 |
| 2 | Type of production in the company | 0.383 |
| 3 | Type of industry | 0.262 |
| 4 | Type of ownership of the company | 0.680 |
| 5 | Company situation | 0.540 |
| 6 | Type of capital in the companies | 0.210 |
| 7 | Types of machines owned | 0.102 |
| 8 | 5S implementing | 0.284 |
| 9 | SMED implementing | 0,001 |
| 10 | Kanban system for spare parts implementing | 0.312 |
| 11 | The way of supervision in the companies | 0.412 |
| 12 | The type of supervision in the companies | 0,000 |
| 13 | The MTTR value | 0.071 |

For the analyzed Hypotheses 9 and 12, there is a statistically difference in the value of the NUD indicator (p-value NUD = 0.001 and NUD = 0.000 - H0 rejected, H1 accepted). It means that there is a statistically justified difference in reducing the NUD from the factors studied. This proves that in the surveyed companies, decreasing the NUD depends on the implementation of the SMED method and from different types of supervision.

The presented analyzes allowed to identify the factors that have impact on the effectiveness of LMn. Moreover, the analyses showed the, which factors did not have the influence on the effectiveness of LMn. Despite the analyzed single factors, for example, such as: types of machines, Kanban, the way of supervision in the companies, it does not have a significant impact on the effectiveness of LMn, their interaction with other factors may already have a significant impact on the LMn effectiveness.

Therefore, in the next stage of the research, the concept of using ML method an RST to search for relationships between the identified factors, and thus their impact on the effectiveness of the LMn concept implementation, was proposed.

4.3. Decision trees in evaluation the effectiveness of Lean Maintenance implementation

Not all surveyed companies used the same LMn tools and methods, therefore CART decision trees were used for analysis. The main criterion for selecting this method was the possibility of its effective use for data sets that have numerous shortcomings in the independent variables. Moreover, this method is insensitive to the occurrence of atypical observations that may come from a different population. The CART classification tree for the dependent variable - reduction in the number of unplanned downtimes (NUD) was developed for the studied group of companies.

In the decision tree the training data set (from 65 companies) and the variables e.g. size of the companies, type of industry, type of production whose impact on the effectiveness of LMn implementation were analyzed (Table 2) as explanatory variables (predictors) were adopted. In addition, the following indicators were introduced: the TPM number of actions indicator (NTPMA), the number of preven-

tive actions indicator (NPA) and the maintenance strategy indicator (MSI). The NPA is the activities number to prevent UD. The NTPMA indicator is calculated as the sum of the value of activities by the maximum number of implemented activities (3):

$$NTPMA = \frac{\sum_{i=1}^{11} x_i}{\max \text{ number of activities}} * 100\% \quad (3)$$

The NTPMA indicator can take values on four levels from low to very high. The calculation of the MSI indicator value is assumed as: the sum of the activities value by the number of implemented activities (4).

$$MSI = \frac{\sum_{i=1}^n x_i}{n} \quad (4)$$

Detailed information about these indicators are presented in the work [4, 5].

While building the tree, the following assumptions were made: the costs of misclassifications were equal, the Gini measure as a measure of goodness, the discontinuation of the process of creating new nodes using trimming according to the variance (the stop rule) and the minimum frequency criterion in the split node, and a 10-fold cross validation as a quality measure. A developed tree consists of 15 divided nodes and 16 end nodes, which means that 16 decision rules may be defined. The developed decision tree is presented on Figure 10.

Selected decision rules were defined for the developed tree. These rules were defined for the end nodes that achieved the best results in reducing NUD using additional LMn methods and tools. Based on the decision tree, the chosen decision rules were defined:

1. If the company's type of supervision expressed by the MSI indicator is different than 5.5, the 5S method is implemented in different areas, it is not a representative of the metal processing industry, it is not a small enterprise and implements a different type of production than small batch production (MS), it achieves a reduction in the NUD in the range from 10 to 30%.
2. If in the enterprise the supervision method expressed by the MSI indicator is different than 5.5, the 5S method is implemented in various areas, it is not a representative of the metal processing industry, the supervision method expressed by the MSI indicator is not equal to 5 or 4, mainly has numerical machines or referred to as "other" machines achieve a reduction in the NUD indicator in the range of 10 to 30%.

3. If in the enterprise the supervision method expressed by the MSI indicator is different than 5.5, the 5S method is implemented in various areas, it is not a representative of the metal processing industry, the supervision method expressed by the MSI indicator is not equal to 5 or 4, mostly it has conventional machines and an average repair time of over 24 hours achieve a reduction in NUD by more than 50%.

In order to evaluate the quality of the developed classification model (DT), the validation for the test data was performed.

The obtained decision rules were used to develop the expert system. For validation the developed decision rules in the expert system the data from 25 companies was used. Among the analyzed companies, the major group were large companies (70%) mainly from the aviation industry (40%). Large batch production dominated (35%) in these companies. Then, using the obtained results the classification quality of the developed decision rules were tested.

The purpose of the qualitative analysis was to generate confusion matrices for the most frequently occurring classes. When developing the confusion matrix, the analyzed class was considered as positive, while other classes were considered as negative. Tables 3 and 4 present confusion matrices for the classifier - the value of NUD for the two the most frequently occurring classes: 10-30% and 30-50%.

Table 3. Confusion matrix for the classifier value of the NUD 10 – 30 % class

| Real Classes | Predicted Classes | |
|--------------|-------------------|----------|
| | Positive | Negative |
| Positive | 11 | 1 |
| Negative | 0 | 13 |

Table 4. Confusion matrix for the classifier value of the NUD 30 - 50 % class

| Real Classes | Predicted Classes | |
|--------------|-------------------|----------|
| | Positive | Negative |
| Positive | 11 | 0 |
| Negative | 2 | 12 |

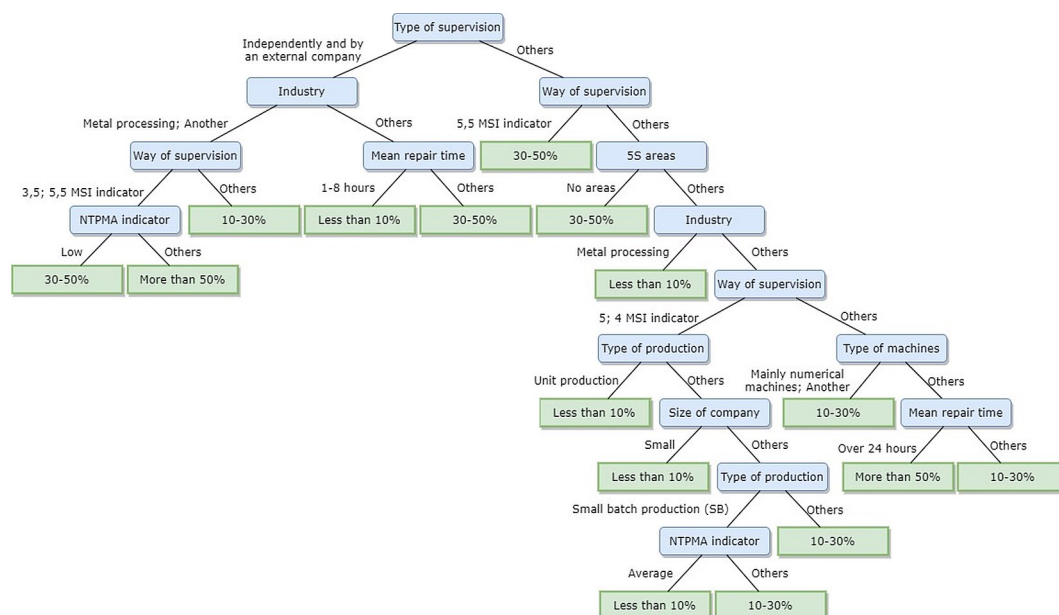


Fig. 10. The developed decision tree

The indicators from Table 2 have been used to assess the quality of the classifier (Table 5).

Table 5. Indicators used to assess the quality of classifier

| Indicators | | | Acc | Err | TPR | TNR | PPV | NPV | FPR | FDR | FNR | MCC | F1 | J |
|-----------------------|--------------|--------|------|------|------|------|------|------|------|------|------|------|------|------|
| Classifier: NUD value | Marked class | 30-50% | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| | | 10-30% | 0.96 | 0.04 | 0.92 | 1.00 | 1.00 | 0.93 | 0.00 | 0.00 | 0.08 | 0.92 | 0.96 | 0.92 |

For easier analysis the results presented in the Table 5, the indicators into two groups were divided. The first (marked in red) contains indicators, of which the value should be as small as possible - in the case of the classifier without errors, the result will be 0. The second of them (other indicators) contains indicators, of which the expected value should be as high as possible = 1. The results presented in the table indicate that the NUD classifier in the 30-50% class is more likely to assign objects to the class to which in fact belong (Acc = 1). For the 10-30% class, the Acc is 0.96, which means that Err = 0.04.

The main goal of the validation was to confirm, that the developed decision rules actually lead to the planned results. The obtained values of calculated indicators confirmed the high usefulness of the classifiers.

4.4. Theory of Rough Sets in Lean Maintenance implementation assessment

In this stage the RST for the described variable NUD was used. In the analyses the same training data set as input (data from 65 companies and explanatory variables (predictors)) were adopted. The following algorithms were used to generate the decisions rules: exhaustive algorithm (ExhAlg), coverage algorithm (CovAlg), genetic algorithm (GenAlg) and LEM2 algorithm. The scheme for the explained variable “reduction in the NUD” is presented on Figure 11.

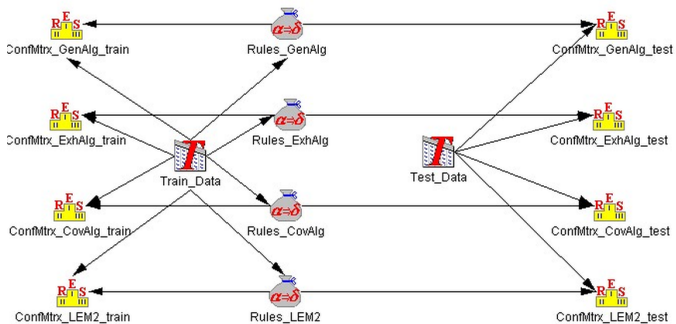


Fig. 11. The scheme for the explained variable “reduction in the NUD”

In Table 6 number of decision rules generated by each algorithm are presented.

Table 6. The number of decisions rules generated for the explained variable NUD

| Name of a Rule Set | Number of Rules |
|--------------------|-----------------|
| NUD_ExhAlg | 6920 |
| NUD_GenAlg | 458 |
| NUD_CovAlg | 43 |
| NUD_LEM2 | 27 |

The rules generated by each algorithm were used to classify the NUD indicator. The classification of objects (companies) from the appropriate decision tables was performed. The standard voting method was used for classification. The results of the classification for each of the algorithms in the form of a confusion matrix is presented. The rows of the matrix show the values for the actual decision classes (the values of the dependent variable). On the other hand, in the columns of the matrix the results of prediction are presented. Additionally, the matrix contains the information about the number of objects belonging to a given decision class, accuracy and coverage. Moreover, a true positive rate is presented.

In the Table 7 the results of classification for GenAlg, ExhAlg and LEM2 are presented. In the case of the explained variable NUD, the confusion matrices were the same for these algorithms. All 65 objects in the decision table were correctly classified (Total Acc = 1).

In the Table 8 and 9 the results of the classification for CovAlg with different value of coverage parameter are presented.

In the case of rules created by the coverage algorithm it was different. When assuming a small value of the coverage equal to 0.001 or less, the algorithm generates rules that give the maximum coverage calculated for all decision classes jointly. It is approximately 0.977 (Table 8). However, with this value of the coverage factor, the classification accuracy is not maximum - it amounts to 0.95. It is caused by an incorrect classification of three objects which have been assigned to the class > 50%. In fact, these objects belong to the decision class of 10-30%. To increase the accuracy of the classification the value of the coverage should be increased. Already for the coverage value equal to 0.12, the accuracy is 1, which means no classification errors (Table 9). However, the coverage is less than that generated previously, and is approximately 0.895. This is due to the lack of classification of two objects from classes <10%, two objects from the class 10 - 30% and one object from the class 30 - 50%.

As in the case of decision trees, the developed decision rules were implemented in the expert system. Again, the data from 25 companies to validate the decision rules was used. To assess the quality of the classifiers the confusion matrices were developed. These confusion matrices by comparison of the results from the studied companies with the result from the expert system were performed. In the Table 10 the results of NUD classification for the LEM2 algorithm are presented. Total Accuracy for this algorithm is 0.958.

In the Table 11 the results of the classification for CovAlg are presented.

Total Accuracy for this algorithm is 0.940, which means that the ability of this classifier is lower than in the case of LEM2 algorithm. In the Table 12 the results of the classification for ExhAlg are presented. Total Accuracy of this classifier is 0.980.

The best results for GenAlg algorithm were obtained. All 25 objects in the decision table were correctly classified (Total Accuracy = 1).

4.5. Results comparison

In the Table 13, the comparison of the results for the most frequently occurring classes: 10–30% 30–50% is presented. The comparison presents the indicators values for the models generated using DT and RST.

Results for the genetic algorithm are not included in the Table 13, because the results are the same as for exhaustive algorithm in the marked class of 10–30%. Considering the 10-30% class, the Accuracy ratio shows that the genetic algorithm and the exhaustive algorithm are most likely to assign objects to the class to which they actually belong. Only a slightly worse Accuracy result was obtained for the

Table 7. Confusion matrix - GenAlg, ExHAlg and LEM2

| Actual | Predicted | | | | | | |
|--------------------|-----------|--------|--------|-------|-------------|----------|----------|
| | < 10% | 10-30% | 30-50% | > 50% | No. of obj. | Accuracy | Coverage |
| < 10% | 44 | 0 | 0 | 0 | 44 | 1 | 1 |
| 10-30% | 0 | 10 | 0 | 0 | 10 | 1 | 1 |
| 30-50% | 0 | 0 | 6 | 0 | 6 | 1 | 1 |
| > 50% | 0 | 0 | 0 | 5 | 5 | 1 | 1 |
| True positive rate | 1 | 1 | 1 | 1 | | | |
| Total Accuracy | 1 | | | | | | |
| Total Coverage | 1 | | | | | | |
| Total no. of obj. | 65 | | | | | | |

Table 8. Confusion matrix - CovAlg (coverage value = 0.001)

| Actual | Predicted | | | | | | |
|--------------------|-----------|--------|--------|-------|-------------|----------|----------|
| | < 10% | 10-30% | 30-50% | > 50% | No. of obj. | Accuracy | Coverage |
| < 10% | 0 | 40 | 0 | 3 | 44 | 0.91 | 0.98 |
| 10-30% | 9 | 0 | 0 | 0 | 10 | 1 | 0.90 |
| 30-50% | 0 | 0 | 6 | 0 | 6 | 1 | 1 |
| > 50% | 0 | 0 | 0 | 5 | 5 | 1 | 1 |
| True positive rate | 1 | 1 | 1 | 0.6 | | | |
| Total Accuracy | 0.977 | | | | | | |
| Total Coverage | 0.95 | | | | | | |
| Total no. of obj. | 65 | | | | | | |

Table 9. Confusion matrix - CovAlg (coverage value = 0.012)

| Actual | Predicted | | | | | | |
|--------------------|-----------|--------|--------|-------|-------------|----------|----------|
| | < 10% | 10-30% | 30-50% | > 50% | No. of obj. | Accuracy | Coverage |
| < 10% | 0 | 42 | 0 | 0 | 44 | 1 | 0.95 |
| 10-30% | 8 | 0 | 0 | 0 | 10 | 1 | 0.80 |
| 30-50% | 0 | 0 | 5 | 0 | 6 | 1 | 0.83 |
| > 50% | 0 | 0 | 0 | 5 | 5 | 1 | 1 |
| True positive rate | 1 | 1 | 1 | 1 | | | |
| Total Accuracy | 1 | | | | | | |
| Total Coverage | 0.895 | | | | | | |
| Total no. of obj. | 65 | | | | | | |

Table 10. Confusion matrix - LEM2

| Actual | Predicted | | | | | | |
|--------------------|-----------|--------|--------|-------|-------------|----------|----------|
| | < 10% | 10-30% | 30-50% | > 50% | No. of obj. | Accuracy | Coverage |
| < 10% | 3 | 0 | 0 | 0 | 3 | 0.917 | 1 |
| 10-30% | 1 | 8 | 0 | 0 | 9 | 0.958 | 1 |
| 30-50% | 0 | 0 | 10 | 0 | 10 | 1.000 | 1 |
| > 50% | 1 | 0 | 0 | 1 | 2 | 0.958 | 1 |
| True positive rate | 0.6 | 1 | 1 | 1 | | | |
| Total Accuracy | 0.958 | | | | | | |
| Total Coverage | 1 | | | | | | |
| Total no. of obj. | 25 | | | | | | |

other two RST algorithms and for DT. The Accuracy results for the 30-50% class are different. The maximum value was obtained for the LEM2 algorithm, the genetic algorithm and for DT. The lowest value was recorded for the coverage algorithm. Similar conclusions can be drawn by looking at the general classifier error (Err) (keeping in mind that the lower the value of the Err, the better the classifier). This shows that the ability to predict of the models created varies depending on the NUD indicator, and the results contained in discussed table may be valuable for future users of the developed models.

The differences in the results can also be seen in cases of sensitivity (TPR), which shows the ability to recognize objects belonging to the distinguished class. For the 30-50% class, the TPR indicator obtained the maximum value for all models except for the classifier generated with the coverage algorithm. However, in the case of the 10-30% class, the LEM2 and DT algorithm did not reach the value of 1. The results of the TPR index are very similar to the NPV, which indicates the probability that an object assigned to the unmarked class by the classifier actually belongs to this class.

One of the best results was obtained for the TNR index, which indicates the ability to correctly classify objects not belonging to the marked class. Comparing the TNR and TPR values for the LEM2 algorithm and DT in the 10-30% class, it can be seen that these classifiers better recognize objects not belonging to this class. A similar situation occurs for the coverage algorithm in the 30-50% class. The values of the Precision index (PPV) were almost identical to those in the TNR.

In the case of the last three indicators from Table 12 (Matthew's correlation coefficient, F1-score, and Youden's J statistic), the results calculated for each of them are similar. All three indicators show that the best classifiers for the marked class 10-30% are classifiers built on the basis of the exhaustive algorithm and the genetic algorithm. However, for the class 30-50%, the best classifiers come from the LEM2 algorithm, the genetic algorithm and DT.

The probability of omitting marked objects by assigning them to an unmarked class is called FNR. This indicator is the lowest in the case of the exhaustive, coverage and genetic algorithms in the 10-30% class. However, in the 30-50% class, all classifiers have the lowest possible FNR value, except for the classifier built on the basis of the coverage algorithm. On the other hand, the FPR and FDR indicators, which refer to the probability of so-called false alarms generated by the classifier, show that the mentioned probability is equal to zero for all classifiers except CovAlg in the 10-30% class, as well as ExhAlg and CovAlg in the 30- class 50%.

Table 11. Confusion matrix - CovAlg (coverage = 0.12)

| Actual | Predicted | | | | | | |
|--------------------|-----------|--------|--------|-------|-------------|----------|----------|
| | < 10% | 10–30% | 30–50% | > 50% | No. of obj. | Accuracy | Coverage |
| < 10% | 3 | 0 | 0 | 0 | 3 | 1.000 | 1 |
| 10–30% | 0 | 9 | 0 | 0 | 9 | 0.960 | 1 |
| 30–50% | 0 | 1 | 9 | 1 | 11 | 0.880 | 1 |
| > 50% | 0 | 0 | 1 | 1 | 2 | 0.920 | 1 |
| True positive rate | 1 | 0.9 | 0.9 | 0.5 | | | |
| Total Accuracy | 0.940 | | | | | | |
| Total Coverage | 1 | | | | | | |
| Total no. of obj. | 25 | | | | | | |

Table 12. Confusion matrix - ExhAlg

| Actual | Predicted | | | | | | |
|--------------------|-----------|--------|--------|-------|-------------|----------|----------|
| | < 10% | 10–30% | 30–50% | > 50% | No. of obj. | Accuracy | Coverage |
| < 10% | 2 | 0 | 1 | 0 | 3 | 0.960 | 1 |
| 10–30% | 0 | 9 | 0 | 0 | 9 | 1.000 | 1 |
| 30–50% | 0 | 0 | 10 | 0 | 10 | 0.960 | 1 |
| > 50% | 0 | 0 | 0 | 3 | 3 | 1.000 | 1 |
| True positive rate | 1 | 1 | 0.91 | 1 | | | |
| Total Accuracy | 0.980 | | | | | | |
| Total Coverage | 1 | | | | | | |
| Total no. of obj. | 25 | | | | | | |

Table 13. Comparison of results – DT and RST

| Indicators | Classifier: reducing the NUD Value | | | | | | | |
|------------|------------------------------------|-------|----------|----------|--------|-------|----------|----------|
| | Marked Class | | | | | | | |
| | 10–30% | | | | 30–50% | | | |
| | DT | RST | | | DT | RST | | |
| | | LEM2 | Exh.Alg. | Cov.Alg. | | LEM2 | Exh.Alg. | Cov.Alg. |
| Acc | 0.960 | 0.958 | 1.000 | 0.960 | 1.000 | 1.000 | 0.960 | 0.880 |
| Err | 0.040 | 0.042 | 0.000 | 0.040 | 0.000 | 0.000 | 0.040 | 0.120 |
| TPR | 0.920 | 0.889 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.818 |
| TNR | 1.000 | 1.000 | 1.000 | 0.938 | 1.000 | 1.000 | 0.933 | 0.929 |
| PPV | 1.000 | 1.000 | 1.000 | 0.900 | 1.000 | 1.000 | 0.909 | 0.900 |
| NPV | 0.930 | 0.938 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.867 |
| FPR | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 | 0.333 |
| FDR | 0.000 | 0.000 | 0.000 | 0.100 | 0.000 | 0.000 | 0.091 | 0.100 |
| FNR | 0.080 | 0.111 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.182 |
| MCC | 0.920 | 0.913 | 1.000 | 0.919 | 1.000 | 1.000 | 0.921 | 0.757 |
| F1 | 0.960 | 0.941 | 1.000 | 0.947 | 1.000 | 1.000 | 0.952 | 0.857 |
| J | 0.920 | 0.889 | 1.000 | 0.938 | 1.000 | 1.000 | 0.933 | 0.747 |

5. Conclusions

Many companies use LM mainly to eliminate production losses. These companies not only increase their productivity, but also strengthen their position on the market. It turns out that companies have started to recognize the importance of maintenance, so they have started implementing LMn.

In this paper the problem of LMn implementation assessment was analyzed. Firstly the data from the manufacturing companies were

collected and preliminary analyzed. The chi-square test for identification the factor affecting for LMn were used.

Then, the machine learning method to develop the classification models was proposed. These models by using DT (CART) and RST (four different algorithms: LEM2, Exh.Alg. Cov.Alg and GenAlg). were developed. To develop these models, data obtained from companies, that implemented LMn were used. In the first stage of the survey, information from companies was collected on: used maintenance strategies, implemented LMn methods and tools, and the results of the implementation. To assess the benefits of the LMn implementation the indicator NUD was analyzed.

The obtained results indicate, that both for the classifiers obtained, RST and DT have a high prediction ability. However, the accuracy of the prediction depends from the analyzed class. The predictive model generated by DT show the better prediction ability in the analyzed class 30–50%. However, the situation in RST is slightly different. The same high prediction ability was demonstrated by the model generated with the use of the genetic algorithm. For the two most frequently occurring classes, this model has the same high predictive ability. However, better accuracy for the class of 30–50% were achieved for RST for LEM2 algorithm. It should be noted that

this algorithm generates the smallest number of decision rules. This shows that a large number of decision rules is not required to obtain good ability of prediction models. For the 10–30% class, the best prediction ability was obtained for the model with the use of the coverage algorithm. The worst prediction ability for the most frequently occurring classes was achieved by models generated with the use of the coverage algorithm.

The created models have some limitations. First of all, these models were developed only based on a small group of companies in

the specific region. Secondly, despite the fact that companies of various sizes and from various industries were invited to participate in the research, large enterprises from the aviation industry were the largest group. As a result, the developed models are based primarily on the experience and effective implementations of LMn by these companies. Therefore, it may be a potential limitation of the implementation of these models in practice. Finally, a high level detailing has been taken to develop the model using DT. This can over-fit the model to the data. Thus, it is planned to continue relevant research in the future to eliminate the limitations of the developed models.

Although the conducted research has some limitations, the presented results can be used by all manufacturing companies to predict and assess the effectiveness of the implementation of LMn methods and tools. In addition, the research results can be used by companies and scientists for the effective organization of maintenance, selection of an appropriate maintenance strategy, but above all for improvement of already implemented activities in this area.

References:

1. Amin MA, Alam MR, Alidrisi H, Karim MA. A fuzzy-based leanness evaluation model for manufacturing organisations. *Production Planning & Control* 2021; 32(11): 959-974, <https://doi.org/10.1080/09537287.2020.1778113>.
2. Antony J, Psomas E, Garza-Reyes JA, Hines P. Practical implications and future research agenda of lean manufacturing: a systematic literature review. *Production Planning & Control* 2020;1-37, <https://doi.org/10.1080/09537287.2020.1776410>.
3. Antosz K, Mazurkiewicz D, Kozłowski E, Sęp J, Żabiński T. Machining Process Time Series Data Analysis with a Decision Support Tool. In: Machado J., Soares F., Trojanowska J., Ottaviano E. (eds) *Innovations in Mechanical Engineering*. iencieng 2021. Lecture Notes in Mechanical Engineering. Springer, Cham. 2021; 14-27, https://doi.org/10.1007/978-3-030-79165-0_2.
4. Antosz K, Paško Ł, Gola A. The use of intelligent systems to support the decision-making process in Lean Maintenance management IFAC PAPERSONLINE 2019; 52(10): 148-153, <https://doi.org/10.1016/j.ifacol.2019.10.037>.
5. Antosz K. *Metodyka modelowania oceny i doskonalenia koncepcji Lean Maintenance*, OW PRZ, Rzeszów, 2019.
6. Arlinghaus JC, Knizkov S. Lean Maintenance and Repair Implementation - A Cross-Case Study of Seven Automotive Service Suppliers. *Procedia CIRP* 2020; 93: 955-964, <https://doi.org/10.1016/j.procir.2020.03.144>.
7. Arrascue-Hernandez G, Cabrera-Brusil J, Chavez-Soriano P, Raymundo-Ibañez C, Perez M. LEAN maintenance model based on change management allowing the reduction of delays in the production line of textile SMEs in Peru. In *IOP Conference Series: Materials Science and Engineering* 2020; 796(1): 012017, <https://doi.org/10.1088/1757-899X/796/1/012017>.
8. Arslankaya S, Atay H. Maintenance management and lean manufacturing practices in a firm which produces dairy products. *Procedia-Social and Behavioral Sciences* 2015; 207: 214-224, <https://doi.org/10.1016/j.sbspro.2015.10.090>.
9. Aucasime-Gonzales P, Tremolada-Cruz S, Chavez-Soriano P, Dominguez F, Raymundo C. Waste Elimination Model Based on Lean Manufacturing and Lean Maintenance to Increase Efficiency in the Manufacturing Industry. In *IOP Conference Series: Materials Science and Engineering* IOP Publisher 2020; 999(1): 012013, <https://doi.org/10.1088/1757-899X/999/1/012013>.
10. Ball P, Lunt P. Lean eco-efficient innovation in operations through the maintenance organisation. *International Journal of Production Economics* 2020; 219: 405-415, <https://doi.org/10.1016/j.ijpe.2018.07.007>.
11. Baluch NH, Che Sobry A, Shahimi M. TPM and Lean maintenance-A Critical Review Interdisciplinary. *Journal of Contemporary Research in Business* 20124 (2): 850-857.
12. Barnard A. Lean Reliability Engineering. In *INCOSE International Symposium* 2014; 24(s1): 13-23, <https://doi.org/10.1002/j.2334-5837.2014.00002.x>.
13. Bar-or A, Schuster A, Wolff R, Keren D. Decision tree induction in high dimensional hierarchically distributed databases. In *Proceedings SI-AM International Data Mining Conference Newport Beach CA* 2005; 466-470, <https://doi.org/10.1137/1.9781611972757.42>.
14. Bertolini M, Mezzogori D, Neroni M, Zammori F. Machine Learning for industrial applications: a comprehensive literature review. *Expert Systems with Applications* 2021; 114820, <https://doi.org/10.1016/j.eswa.2021.114820>.
15. Bhasin S. Prominent obstacles to lean. *International Journal of Productivity and Performance Management* 2011; 61(4): 403-425, <https://doi.org/10.1108/17410401211212661>.
16. Bhuvanesh Kumar M, Parameshwaran R. Fuzzy integrated QFD FMEA framework for the selection of lean tools in a manufacturing organisation. *Production Planning & Control* 2018; 29(5): 403-417, <https://doi.org/10.1080/09537287.2018.1434253>.
17. Bortolotti T, Boscari S, Danese P. Successful lean implementation: Organizational culture and soft lean practices. *International Journal of Production Economics* 2015; 160: 182-201, <https://doi.org/10.1016/j.ijpe.2014.10.013>.
18. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall New York 1984.
19. Bukowski L, Werbińska-Wojciechowska S. Using fuzzy logic to support maintenance decisions according to Resilience-Based Maintenance concept. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2021; 23 (2): 294-307, <https://doi.org/10.17531/ein.2021.2.9>.
20. Campagner A, Ciucci D, Hüllermeier E. Rough set-based feature selection for weakly labeled data. *International Journal of Approximate Reasoning* 2021; 136: 150-167, <https://doi.org/10.1016/j.ijar.2021.06.005>.
21. Ceruti A, Marzocca P, Liverani A, Bil C. Maintenance in aeronautics in an industry 4.0 context: the role of augmented reality and additive manufacturing. *Journal of Computational Design and Engineering* 2019; 6(4): 516-526, <https://doi.org/10.1016/j.jcde.2019.02.001>.
22. Chemweno P, Pintelon L, Muchiri PN, Van Horenbeek A. Risk assessment methodologies in maintenance decision making: A review of dependability modelling approaches. *Reliability Engineering & System Safety* 2018; 173: 64-77, <https://doi.org/10.1016/j.res.2018.01.011>.
23. Chen C, Wang C, Lu N, Jiang B, Xing Y. A data-driven predictive maintenance strategy based on accurate failure prognostics. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2021; 23 (2): 387-394, <https://doi.org/10.17531/ein.2021.2.19>.
24. Chowdary BV, Ojha K, Alexander A. Improvement of refinery maintenance and mechanical services: application of lean manufacturing principles. *International Journal of Collaborative Enterprise* 2018; 6(1): 20-36, <https://doi.org/10.1504/IJCEN.2018.092082>.
25. Ciano MP, Pozzi R, Rossi T, Strozzi F. How IJPR has Addressed 'Lean': A Literature Review Using Bibliometric Tools. *International Journal of Production Research* 2019; 57 (15-16): 5284-5317, <https://doi.org/10.1080/00207543.2019.1566667>.
26. Damián M, Chambilla M, Viacava G, Eyzaguirre J, Raymundo C. Lean Service Model for Maintenance Management Using a Linear Programming Approach. In *2021 10th International Conference on Industrial Technology and Management (ICITM) 2021*; 25-30, <https://doi.org/10.1109/ICITM52822.2021.00012>.

27. Dekker R. Applications of maintenance optimisation models: A review and analysis. *Reliability Engineering and System Safety* 1996; 51: 229-240, [https://doi.org/10.1016/0951-8320\(95\)00076-3](https://doi.org/10.1016/0951-8320(95)00076-3).
28. Drożnyer P. The impact of the implementation of management system on the perception of role and tasks of maintenance services and effectiveness of their functioning. *Journal of Quality in Maintenance Engineering* 2021; 27(2): 430-450, <https://doi.org/10.1108/JQME-09-2019-0089>.
29. Duran O, Capaldo A, Acevedo PAD. Lean maintenance applied to improve maintenance efficiency in thermoelectric power plants. *Energies* 2017; 10(10): 1-22, <https://doi.org/10.3390/en10101653>.
30. Epler I, Sokolović V, Milenkov M, Bukvić M. Application of lean tools for improved effectiveness in maintenance of technical systems for special purposes. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2017; 19 (4): 615–623, <https://doi.org/10.17531/ein.2017.4.16>.
31. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006; 27: 861-874, <https://doi.org/10.1016/j.patrec.2005.10.010>.
32. Gao R, Wang L, Teti R, Dornfeld D, Kumara S, Mori M, Helu M. Cloud-enabled Prognosis for Manufacturing. *CIRP Annals-Manufacturing Technology* 2015; 64(2): 749-772, <https://doi.org/10.1016/j.cirp.2015.05.011>.
33. Gaur J, Goel AK, Rose A, Bhushan B. Emerging trends in machine learning. In 2019 2nd International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICT) IEEE 2019; 1: 881-885, <https://doi.org/10.1109/ICICT46008.2019.8993192>.
34. Godara S, Singh R. Evaluation of predictive machine learning techniques as expert systems in medical diagnosis. *Indian Journal of Science and Technology* 2016; 9(10): 1-14, <https://doi.org/10.17485/ijst/2016/v9i10/87212>.
35. Gupta S, Gupta P, Parida A. Modeling lean maintenance metric using incidence matrix approach. *International Journal of System Assurance Engineering and Management* 2017; 8(4): 799-816, <https://doi.org/10.1007/s13198-017-0671-z>.
36. Henríquez-Alvarado F, Luque-Ojeda V, Macassi-Jauregui I, Alvarez JM, Raymundo-Ibañez C. Process optimization using lean manufacturing to reduce downtime: Case study of a manufacturing SME in Peru. Paper presented at the ACM International Conference Proceeding Series 2019; 261-265.
37. Holgado M, Macchi M, Evans S. Exploring the impacts and contributions of maintenance function for sustainable manufacturing. *International Journal of Production Research* 2020; 58(23): 7292-7310, <https://doi.org/10.1080/00207543.2020.1808257>.
38. Jasiulewicz-Kaczmarek M, Antosz K, Wyczółkowski R, Mazurkiewicz D, Sun B, Qian C, Ren Y. Application of MICMAC, Fuzzy AHP, and Fuzzy TOPSIS for Evaluation of the Maintenance Factors Affecting Sustainable Manufacturing. *Energies* 2021; 14(5): 1436, <https://doi.org/10.3390/en14051436>.
39. Jasiulewicz-Kaczmarek M, Antosz K, Zywicka P, Mazurkiewicz D, Sun B, Ren Y. Framework of machine criticality assessment with criteria interactions. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2021; 23(2): 207-220, <https://doi.org/10.17531/ein.2021.2.1>.
40. Jasiulewicz-Kaczmarek M, Saniuk A. How to Make Maintenance Processes More Efficient Using Lean Tools?. In *International Conference on Applied Human Factors and Ergonomics*, Cham 2018; 9-20, https://doi.org/10.1007/978-3-319-60828-0_2.
41. Jasiulewicz-Kaczmarek M, Żywicka P, Gola A. Fuzzy set theory driven maintenance sustainability performance assessment model: A multiple criteria approach. *Journal of Intelligent Manufacturing* 2021; 32(5): 1497-1515, <https://doi.org/10.1007/s10845-020-01734-3>.
42. Jing S, Niu Z, Chang PC. The application of VIKOR for the tool selection in lean management. *Journal of Intelligent Manufacturing* 2019; 30(8): 2901-2912, <https://doi.org/10.1007/s10845-015-1152-3>.
43. Kammerer K, Hoppenstedt B, Pryss R, Stokler S, Allgaier J, Reichert M. Anomaly detections for manufacturing systems based on sensor data-insights into two challenging real-world production settings. *Sensors* 2019; 19(24): 5370, <https://doi.org/10.3390/s19245370>.
44. Khazravi N, Alavi SM. A New Method To Feature Selection In Rough Fuzzy Set Theory Based On Degree Of Separation Turkish. *Journal of Computer and Mathematics Education (TURCOMAT)* 2021; 12(14): 1889-1897.
45. Kovács G. Combination of Lean Value-Oriented Conception and Facility Layout Design for Even More Significant Efficiency Improvement and Cost Reduction. *International Journal of Production Research* 2020; 58 (10): 2916-2936, <https://doi.org/10.1080/00207543.2020.1712490>.
46. Kuhnle A, Jakubik J, Lanza G. Reinforcement learning for opportunistic maintenance optimization. *Production Engineering* 2018; 13(1): 33-41, <https://doi.org/10.1007/s11740-018-0855-7>.
47. Kumar MB, Parameshwaran R. A comprehensive model to prioritise lean tools for manufacturing industries: a fuzzy FMEA AHP and QFD-based approach. *International Journal of Services and Operations Management* 2020; 37(2): 170-196, <https://doi.org/10.1504/IJSOM.2020.110337>.
48. Larose DT. Discovering knowledge from data Introduction to data mining. Scientific publisher PWN Warsaw 2013, <https://doi.org/10.1002/9781118874059>.
49. Lundgren C, Skoogh A, Bokrantz J. Quantifying the effects of maintenance-a literature review of maintenance models. *Procedia CIRP* 2018; 72: 1305-1310, <https://doi.org/10.1016/j.procir.2018.03.175>.
50. Ma J, Atef M, Nada S, Nawar A. Certain Types of Covering-Based Multigranulation -Fuzzy Rough Sets with Application to Decision-Making Complexity 2020; Article ID 6661782, <https://doi.org/10.1155/2020/6661782>.
51. Macchi M, Roda I, Fumagalli L. On the Advancement of Maintenance Management Towards Smart Maintenance in Manufacturing. *IFIP International Conference on Advances in Production Management Systems (APMS) Hamburg Germany Springer, Cham.* 2017; 383-390, https://doi.org/10.1007/978-3-319-66923-6_45.
52. Marksberry P. The Modern Theory of the Toyota production System: A Systems Inquiry of the world's most emulated and profitable management system. CSR Press, Taylor & Francis Group, New York 2013.
53. Marttonen-Arola S, Baglee D, Kinnunen SK, Holgado M. Introducing Lean into Maintenance Data Management: A Decision Making Approach. In: Liyanage J Amadi-Echendu J Mathew J (eds) *Engineering Assets and Public Infrastructures in the Age of Digitalization Lecture Notes in Mechanical Engineering Springer Cham* 2020; https://doi.org/10.1007/978-3-030-48021-9_28.
54. Matyas K, Nemeth T, Kovacs K, Glawar R. A procedural approach for realizing prescriptive maintenance planning in manufacturing industries. *CIRP Annals* 2017; 66(1): 461-464, <https://doi.org/10.1016/j.cirp.2017.04.007>.
55. McCarthy D, Rich N. *Lean TPM A Blueprint for Change*. Elsevier Butterworth-Heinemann 2004, <https://doi.org/10.1016/B978-075065857-7/50005-6>.
56. Mosyurchak A, Veselkov V, Turygin A, Hammer M. Prognosis of behaviour of machine tool spindles their diagnostics and maintenance. *MM Science Journal* 2017; 2100-2104, https://doi.org/10.17973/MMSJ.2017_12_201794.
57. Mouzani IA, Bouami DRISS. The integration of lean manufacturing and lean maintenance to improve production efficiency. *International*

- Journal of Mechanical and Production Engineering Research and Development 2019; 9(1): 601-612, <https://doi.org/10.24247/ijmperdfeb201957>.
58. Muchiri P, Pintelon L, Gelders L, Martin H. Development of maintenance function performance measurement framework and indicators. *International Journal of Production Economics* 2011; 131(1): 295-302, <https://doi.org/10.1016/j.ijpe.2010.04.039>.
 59. Mumani AA, Magableh GM, Mistarihi MZ. Decision making process in lean assessment and implementation: a review. *Management Review Quarterly* 2021; 1-40, <https://doi.org/10.1007/s11301-021-00222-z>.
 60. Ndhafef N, Nidhal R, Hajji A, Bistorin O. Environmental issue in an integrated production and maintenance control of unreliable manufacturing/remanufacturing systems. *International Journal of Production Research* 2020; 58(14): 4182-4200, <https://doi.org/10.1080/00207543.2019.1650212>.
 61. Nowotarski P, Paślowski J, Dallasega P. Multi-Criteria Assessment of Lean Management Tools Selection in Construction. *Archives of Civil Engineering* 2021; 711-726.
 62. Pasko Ł, Setlak G. Badanie jakości predykcijnej segmentacji rynku. *Zeszyty Naukowe Politech Śląskiej Seria Informatyka* 2016; 37: 83-97.
 63. Pawlak Z, Polkowski L, Skowron A. Rough set theory. *KI* 2001; 15(3): 38-39, https://doi.org/10.1007/978-3-7908-1776-8_1.
 64. Pawlak Z. *Rough Sets: Theoretical Aspects of Reasoning about Data* Theory and Decision. Library D: Springer Netherlands 1991, https://doi.org/10.1007/978-94-011-3534-4_7.
 65. Peres RS, Barata J, Leitao P, Garcia G. Multistage quality control using machine learning in the automotive industry. *IEEE Access* 2019; 7: 79908-79916, <https://doi.org/10.1109/ACCESS.2019.2923405>.
 66. Phogat S, Gupta AK. Theoretical analysis of JIT elements for implementation in the maintenance sector of Indian industries. *International Journal of Productivity and Quality Management* 2018; 25(2): 212-224, <https://doi.org/10.1504/IJPQM.2018.094765>.
 67. Pinto GFL, Silva FJG, Campilho RDSG, Casais R B, Fernandes A J, Baptista A. Continuous improvement in maintenance: a case study in the automotive industry involving Lean tools. *Procedia Manufacturing* 2019; 38: 1582-1591, <https://doi.org/10.1016/j.promfg.2020.01.127>.
 68. Pombal J, Ferreira LP, Sá J C, Pereira MT, Silva FJG. Implementation of lean methodologies in the management of consumable materials in the maintenance workshops of an industrial company. *Procedia Manufacturing* 2019; 38: 975-982, <https://doi.org/10.1016/j.promfg.2020.01.181>.
 69. Qu J, Bai X, Gu J, Taghizadeh-Hesary F, Lin J. Assessment of Rough Set Theory in Relation to Risks Regarding Hydraulic Engineering Investment Decisions. *Mathematics* 2020; 8(8): 1308, <https://doi.org/10.3390/math8081308>.
 70. Ramos E, Mesia R, Alva C, Miyashiro R. Applying lean maintenance to optimize manufacturing processes in the supply chain: A Peruvian print company case. *International Journal of Supply Chain Management* 2020; 9: 264-281.
 71. Ravikumar S, Ramachandran KI, Sugumaran V. Machine learning approach for automated visual inspection of machine components. *Expert systems with applications* 2011; 38(4): 3260-3266, <https://doi.org/10.1016/j.eswa.2010.09.012>.
 72. Rødseth H, Schjølberg P. Data-driven predictive maintenance for green manufacturing In *Proceedings of the 6th international workshop of advanced manufacturing and automation. Advances in Economics Business and Management Research* Atlantis Press 2016; 36-41.
 73. Sakthi Nagaraj T, Jeyapaul R, Vimal KEK, Mathiyazhagan K. Integration of human factors and ergonomics into lean implementation: ergonomic-value stream map approach in the textile industry. *Production Planning & Control* 2019; 30(15): 1265-1282, <https://doi.org/10.1080/09537287.2019.1612109>.
 74. Saxena A, Saad A. Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems. *Applied Soft Computing* 2007; 7(1): 441-454, <https://doi.org/10.1016/j.asoc.2005.10.001>.
 75. Shanmuganathan VK, Haran AP, Gayathri N. Condition monitoring maintenance of aero-engines through LUMS-A method for the implementation of Lean tools. *Measurement* 2015; 73: 226-230, <https://doi.org/10.1016/j.measurement.2015.05.017>.
 76. Shou W, Wang J, Wu P, Wang X. Lean management framework for improving maintenance operation: development and application in the oil and gas industry. *Production Planning & Control* 2021; 32(7):585-602, <https://doi.org/10.1080/09537287.2020.1744762>.
 77. Shou W, Wang J, Wu P, Wang X. Value adding and non-value adding activities in turnaround maintenance process: classification validation and benefits. *Production Planning & Control* 2020; 31(1): 60-77, <https://doi.org/10.1080/09537287.2019.1629038>.
 78. Sidhu SS, Singh K, Ahuja IS. An empirical investigation of maintenance practices for enhancing manufacturing performance in small and medium enterprises of northern India. *Journal of Science and Technology Policy Management* 2021; <https://doi.org/10.1108/JSTPM-11-2019-0109>.
 79. Simões JM, Gomes CF, Yasin MM. Changing role of maintenance in business organisations: measurement versus strategic orientation. *International Journal of Production Research* 2016; 54(11): 3329-3346, <https://doi.org/10.1080/00207543.2015.1106611>.
 80. Singh AK, Vinodh S, Vimal KEK. Application of Grey based decision making approach for lean tool selection. In *5th International & 26th All India Manufacturing Technology Design and Research Conference (AIMTDR 2014) December 12th-14th 2014 IIT Guwahati Assam India* 2014.
 81. Skowron A, Dutta S. Rough sets: past present and future. *Natural computing* 2018; 17(4): 855-876, <https://doi.org/10.1007/s11047-018-9700-3>.
 82. Smith R, Hawkins B. *Lean maintenance; reduce cost improve quality and increase market share*. Elsevier Butterworth-Heinemann 2004.
 83. Sokolova M, Lapalme GA. Systematic analysis of performance measures for classification tasks. *Information Process Management* 2009; 45: 427-437, <https://doi.org/10.1016/j.ipm.2009.03.002>.
 84. Świdorski A, Borucka A, Grzelak M, Gil L. Evaluation of Machinery Readiness Using Semi-Markov Processes. *Applied Sciences* 2020; 10(4):1541, <https://doi.org/10.3390/app10041541>.
 85. Thawkar A, Tambe P, Deshpande V. A reliability centred maintenance approach for assessing the impact of maintenance for availability improvement of carding machine. *International Journal of Process Management and Benchmarking* 2018; 8(3): 318-339, <https://doi.org/10.1504/IJPMB.2018.092891>.
 86. Tortorella GL, Fogliatto FS, Cauchick-Miguel PA, Kurnia S, Jurburg D. Integration of Industry 4.0 technologies into Total Productive Maintenance practices. *International Journal of Production Economics* 2021; 240: 108224, <https://doi.org/10.1016/j.ijpe.2021.108224>.
 87. Traini E, Bruno G, D'antonio G, Lombardi F. Machine learning framework for predictive maintenance in milling. *IFAC-PapersOnLine* 2019; 52(13): 177-182, <https://doi.org/10.1016/j.ifacol.2019.11.172>.
 88. Van Horenbeek A, Kellens K, Pintelon L, Duflou JR. Economic and environmental aware maintenance optimization. *Procedia CIRP* 2014,

- 15: 343-348, <https://doi.org/10.1016/j.procir.2014.06.048>.
89. Velmurugan RS, Dhingra T. Maintenance Strategy Selection and its Impact in Maintenance Function: a Conceptual Framework. *International Journal of Operations & Production Management* 2015; 35 (12): 1622-1661, <https://doi.org/10.1108/IJOPM-01-2014-0028>.
90. Welte R, Estler M, Lucke D. A Method for Implementation of Machine Learning Solutions for Predictive Maintenance in Small and Medium Sized Enterprises. *Procedia CIRP* 2020; 93: 909-914, <https://doi.org/10.1016/j.procir.2020.04.052>.
91. Wu Z, Xu J, Xu Z. A multiple attribute group decision making framework for the evaluation of lean practices at logistics distribution centres. *Annals of Operations Research* 2016; 247(2): 735-757, <https://doi.org/10.1007/s10479-015-1788-6>.
92. Ylipää T, Skoogh A, Bokrantz J, Gopalakrishnan M. Identification of maintenance improvement potential using OEE assessment. *International Journal of Productivity and Performance Management* 2017; 66(1): 126-143, <https://doi.org/10.1108/IJPPM-01-2016-0028>.
93. Zhang C, Wang C, Chen Q. Design of Lean Maintenance Process for Ball Screw Actuator. In 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT) IEEE 2019; 425-430, <https://doi.org/10.1109/ICCASIT48058.2019.8973230>.

The study on the automated storage and retrieval system dependability

Indexed by:



Konrad Lewczuk^a

^aWarsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland


Highlights

- The problem of ARS dependability is discussed for the first time.
- New contribution to discussion on logistics systems dependability is provided.
- New factors influencing the dependability of ASRS are gathered.
- A new simulation model in FlexSim for ASRS dependability analyse is provided.

Abstract

Automated storage systems have become the basis of warehouse logistics. The article presents a discussion on the reliability and dependability of Automated Storage and Retrieval Systems (ASRS), which are perceived as solutions with high technical reliability. Still, their role in the dependability of the entire warehouse system is to be discussed. The concepts of reliability and dependability in logistics systems like ASRS are defined, and a literature review in this area is presented. On this basis, the factors influencing the dependability of ASRS are discussed in a way not present in the discussion on this topic so far. Then, the ASRS simulation model (based on FlexSim simulation software) is presented. The model tests the influence of ASRS configuration and assigned resources on the dependability of the warehouse as a master system. The summary includes observations on defining the reliability and dependability of ASRS.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

Automated Storage and Retrieval System, ASRS, niezawodność, FlexSim, symulacja, magazynowanie.

1. Introduction

Automated storage and retrieval systems (ASRS) are the key components of automated warehouse facilities of high throughput and storage capacity. ASRSs are the automatic solutions around which the warehouse process is built-in or which directly feed high-efficient order-picking or production systems. ASRS defines the physical aspects of the facility, is often an integral part of the picking system, and creates the buffer capacity of the warehouse. The spectrum of ASRS technological solutions and variants is vast. However, the set of common features and mechanisms can be distinguished and put into its definition. ASRSs revolutionize warehousing since the 1970s. One of the most important features of these systems deciding about its usability is the reliability or dependability of this technology.

The system's reliability is a component of its dependability, defined as the ability to perform as and when required [17]. Dependability is then a holistic measure of availability, reliability, maintainability, and maintenance support provided. In some cases, it covers durability, safety, and security [17] to describe how users can trust the services within a time period. Since the ASRS is not an isolated system but a part of the warehousing facility, it should be discussed in the broad context of its dependability (see section 3 for discussion on dependability). In contrast, its dependability is not researched, while reliability research is scarce.

ASRSs are perceived dependable, especially when appropriate maintenance is provided, the system is well configured, and support-

ed by solutions that guarantee the high quality of handled units [24, 40, 48]. But the perception about the dependability of ASRS is a bit warped by the users and developers. In most cases, it refers to the system's uptime (see [24]) and downtime as it results from the definition of reliability. Still, when investigated deeper, the ASRS reliability (or dependability) is rarely explored and usually replaced in the literature and commercial offers by the performance. Performance determines the ability to perform the logistics tasks of the entire ASRS. Usually, it is assumed that it is not significantly affected by the failure of a particular system component so that dependability can be (to some extent) extrapolated by performance bypassing the engineering correctness. Replacing the dependability with the performance requires (or allows for) significant simplifications in the research and development, and most important – in selling. When dependability is removed, the performance is easy to measure. But replacing dependability with performance features requires a set of simplifying assumptions that the material flow in ASRS is uniform and homogeneous (without family grouping or selectivity), no slotting mechanisms are used, and the access to all rack aisles is not disturbed by failures or congestion on feeding conveyor system. With this simplified approach, it is possible to express the transition of ASRS to a state of partial unfitness through reduced performance. This approach is applicable only in general considerations, but applied to the operational level can result in process errors in advanced storage systems. These errors will be the result of limited access to selected product or family groups in the

E-mail addresses: K. Lewczuk - Konrad.lewczuk@pw.edu.pl

ASRS, which will disturb the schedule of the warehouse process and shipments (cf. Kłodawski et al. [23] and Jacyna-Golda et al. [21]).

A review of market solutions and currently published scientific studies on the Automated Storage and Retrieval Systems indicates that ASRS dependability is discussed in a simplified manner or only concerns technical functioning. Meanwhile, ASRSs are built into the warehouse process. Their functioning depends not only on their features but also on external conditions imposed by the supplies, order picking organization, schedule, and shipments. The assessment of ASRS dependability needs to consider the organization of the material flows resulting in an uneven load on the ASRS components, their role in the warehouse process, and the effects of a potential shutdown.

The author proposes the simulation study of the dependability of typical ASRS system in the FlexSim environment. The study bases not only on the technical features of the system but also its configuration (different number of corridors to stacker cranes, conveyor system, transferring module), potential damage to working and conveying elements, the condition of material units, warehouse activity profiling, and assortment distribution. The research will determine the impact of the above-mentioned factors on the dependability and performance of ASRS and the value of OTIFEF (on-time, in-full, error-free parameter) embracing logistics time measures and timely execution of warehouse tasks.

The remainder of this article is as follows. Section 2. provides the literature review on the ASRS reliability or dependability in the context of warehouse processes. The 3. section discusses the problem of dependability in logistics systems and in ASRSs and the areas of dependability that should be investigated for a complete picture of the problem. The 4. section contains the assumptions and measures for the simulation experiments described in the section 5. Section 6. is for ASRS dependability simulation and discussion of the results. The article is closed by the discussion of results and conclusions.

2. Literature review on dependability of automated storage and retrieval systems

The literature on automated storage and retrieval systems is quite extensive and multi-threaded due to the great importance of these solutions for automated warehouse facilities, but reliability and especially dependability of ASRS as complete systems have hardly been studied in academic literature. The literature can be divided into several cross-sectional research areas, within which analytical, experimental, implementation and review works are present. Despite the fact that reliability is a key technical parameter of industrial installations, this issue is not a popular topic of research in warehousing technology at all. Most of the publications on ASRS, usually in the Introduction section, indicate the high reliability of these solutions, but apart from a single word at the beginning, it is not referred to further.

Nowakowski [36, 38, 37], Werbińska-Wojciechowska [46, 47], Bukowski and Feliks [8, 7], or Quigley and Walls [39] present general considerations on the reliability of logistics systems and complex supply chains on the overall level. These publications provide a certain basis for defining the reliability of elements of logistics systems, including warehouse systems, but are not focused on details of technology and technical solutions. The advantage of these studies is the consolidation of the ASRS dependability issue in the science of the reliability of logistics systems. Numerous studies related to problems of dependability in logistics [3, 4, 21] are focused on the reduced efficiency of the system. Sohn and Choi [43] analyse issues related to managing a supply chain in relation to the reliability of subsequent stages – logistic processes, including warehouse processes. They emphasise the need to include reliability issues already at the stage of designing, but their considerations are on the general level. Jacyna and Semenov [19] discuss the topic from the perspective of information uncertainty. Szaciłło et al. [44] touch the problems of reliability applied to railway systems.

The crucial feature of dependability of supply chain, warehouse or ASRS itself is the determination of the faultless probability [22]. This is difficult for structures like ASRSs, in the case of which classical damage causing lack of fitness of use is not applicable.

The problem of the dependability of warehouse facilities and their elements is discussed in general by Lewczuk [27] and Jacyna and Lewczuk [22]. They define the reliability framework for warehouse facilities and their components that may be useful for the assessment of ASRS systems. The authors discuss the OTIFEF index (on-time, in-full, error-free) that can be used to evaluate ASRS similarly as to complete warehouse since this bodies have common definition points. Neo et al. [34] analyse how the limited warehouse technical efficiency influences criteria of its operation assessment. Werbińska-Wojciechowska [47] presents a model of maintaining technical systems on the example of logistic systems using the concept of time delays. Author points to the effectiveness of the devised model on the example of internal transport devices. In other work Werbińska-Wojciechowska [46] discusses the integration of the system executing the task with the supportive system like the maintenance system.

Focusing on the problem of Automated Storage and Retrieval Systems can already see that it reached the cross-sectional publications presenting the state of knowledge about it. Roodbergen and Vis [41], Gagliardi, Renaud, and Ruiz [14], and Azadeh et al. [2] provided a comprehensive literature review on automatic technologies in warehousing. Still, the reliability is mentioned only without discussion, while the dependability is not mentioned at all. Marchet et al. [32] propose a framework for developing and designing some versions of automated storage and retrieval systems but address mostly the system performance and don't mention the reliability or dependability. The majority of publications deal indirectly with the ASRS dependability and its components. The situation when the high reliability of ASRS is called by the authors in the introduction but never referred to in the text is common and applies to all listed publications. This is typical for research on ASRS, which focuses mostly on time efficiency and performance.

Two important studies relating directly to the reliability of the ASRS were conducted by the Material Handling Industry of America and reported by Kluwec in a White Paper: Reliability of Automated Storage/Retrieval Systems (ASRS) [24]. Studies investigated systems in size from 1 to 25 aisles, with an average size being 7.4 aisles (57% of systems had only 1 to 5 aisles). Both studies confirmed the expected high reliability of these solutions, taking the uncertainty out of a long-standing question about ASRS performance. The top concerns of users formulated in the White Paper are downtime (unreliability), potential low flexibility, sunk costs, customer service, implementation, and maintenance issues. Perceptions of low reliability may be related to the experience during the trying period (about three months), even though new ASRS in most cases have fairly high uptime and full performance gain within the first year. The survey shows that uptime increases insignificantly in the first year of operation from 94.05% to 96.22%, and after ten years of operation is decreasing, but still not significantly. The average uptime for the group of respondents was 97.34% during the full performance period [24].

The White Paper [24] reports that insertion/extraction equipment posed the greatest problems for almost 40% of respondents, and the control software was in second place but only for the first three years. The report shows that fast recovery is crucial for minimizing downtime. To that end, the warehouse must have quick access to skilled personnel and immediate availability of needed repair parts. The scheduled maintenance did not have a significant impact on overall uptime while the majority of system downtime was unexpected.

An important factor of ASRS reliability is presented by Ripple [40], which calls pallet/load condition a cause of ASRS faults. These faults are excluded from availability calculations, similarly to the time between the fault occurrence and addressing the problem by personnel. Ripple concludes that equipment failures are quite rare, and when

totes or high-quality pallets are used, the reliability rate can exceed 99.98% or 99.99%.

Methods aimed at researching and increasing the reliability of warehouses use a variety of techniques. Chung, Chan and Chan [9] propose genetic algorithms for maximizing handling reliability of distribution centers. Fazlollahtabar and Saidi-Mehrabad [12] use multi-objective methods for assessing reliability of AGV systems in a multiple AGV jobshop manufacturing system with fuzzy logic. The methods are applied to ASRS exactly as tool presented by Jacyna, Wasiak and Bobiński [19] and Jachimowski et al. [18]. The tool for integrated modelling and simulation of material handling and storage solutions can simulate ASRS in warehouse and state its reliability-related parameters present in the databases for the tool.

An interesting approach to modelling of reliability of warehouse automatic systems was presented by Yan, Dunnett, and Jackson [49], who investigated the reliability of automated guided vehicles system through Failure Modes Effects and Criticality Analysis and then the Fault Tree Analysis (FTA) to model the causes of phase failure. The authors focus on mechanical and constructional aspects of the system and not on the organization or surroundings influence, but their approach can be developed with these factors.

Yang et al. [50] research the problem of goods location assignment in automatic warehouses. Ekren et al. [11] add the element of class-based storage policy to automatic storage and retrieval systems. Both studies prove that proper assignment function for optimizing the cargo space and optimizing the stacker crane operation route can improve overall operating efficiency, which is a part of uptime rationalization. A similar problem, but formulated concerning order-picking, is presented by Atmaca and Ozturk [1]. They show that the appropriate storage assignment in ASRS impacts picking efficiency, thus discussing dependability of ASRS fragmentarily as an element of a larger system. The class-based storage allocation in ASRS was also the main thread of work [30] by Manzini, Gamberi, and Regattieri. Their multi-parametric dynamic model of a product-to-picker assignment and simulation tool confirmed that ASRS should be considered an important chain in the warehousing process.

Liu et al. [28] represent a wide group of authors focusing on travel time models for different automated storage and retrieval systems versions which are important for reliability assessment. Liu et al. provided an extended comparison of models present in the literature and look for better system efficiency, which forms performance characteristics and influences the reliability expressed through uptime function. The models are not very different than those presented by Sarker and Babu [42] in 1995. Boysen and Stephan [5] present a survey on scheduling methods applied to ASRS cranes work organization, like the one presented by Hachemi and Besombes [15] or Zhang et al. [51]. Different approaches are used in these papers, like statistical modelling [44], analytical modelling [31, 26], simulation [25, 10, 35], model predictive control [33], and optimization of all types [13, 50] including evolution algorithms [6]. These studies aim to model and optimise ASRS cycle time, a base for performance analysis, and touch on the problem of material assignment and its influence on the operation. Authors combine elements of spatial configuration, handling equipment, task interleaving, and material assignment but do not touch the dependability issues.

The literature review showed that the reliability and dependability of ASRS are not raised in the literature. This may be the extent and multifactorial nature of this problem and the inability to indicate unambiguous guidelines regarding the dependability, which depends on several factors external to ASRS. The literature does not discuss the impact of the configuration of ASRS racks and conveyors on dependability and the impact of material assignment or tasks resulting from customer orders.

3. The aspects of ASRS dependability

3.1. Dependability of logistics systems

The reliability of the systems is a component of its *dependability* as it results from the definition presented in [17]. This is especially important for logistics systems like ASRS. ASRS is considered a logistics system since it has the buffering capacity, material handling components to transform the material flow, and input and output defined by the qualitative and quantitative material flow structures. In consequence, dependability is a better way to describe its global features than the commonly used reliability. Dependability is a set of features, including readiness, reliability, maintainability, and maintenance support for the system [17, 22]. Nowakowski [38] defines the dependability of any logistics system as a measure of task implementation over time, which may be compared to the reliability of the technical system. He states that no equivalent of maintainability or reliability of the technical system has been formulated for logistics systems of large scale. Still, both terms can be applied to the ASRS when the assumptions are made, especially in a colloquial sense. Nowakowski also defines the dependability of the system through its availability. In technical science, the availability of a recoverable object describes the probability of its proper functioning in a specific moment [22]. Still, the ASRS's availability can be defined as the probability of finding a piece of equipment at any given time during the period of operation, in a state which will allow a requested operation to be carried out correctly and without malfunction [40]. It depends on the availability of resources; cranes, transfers, conveyors, empty storage locations, or required material in the rack (see [46] and *Logistics Management Institute* definitions).

Dependability is a factor difficult to measure considered in designing logistic and warehouse systems. It can be indirectly measured by the disturbances and reduction of the system's performance [16, 37]. The additional measurements are created to reflect the flexibility of the system – its ability to adapt and overcome the difficulties [22], which can be interpreted as the possibility to reconfigure or use other pieces of the system to bypass those unavailable or damaged for task completion.

3.2. Dependability of ASRS

The dependability of ASRS is briefly discussed in the literature, and, as the literature query shows, it is also not an element of the material handling systems design procedure. Both the designers of automatic solutions and a few scientific works refer to the reliability of ASRS, which is based on the failure rate of technical devices that make up the system. Since such a failure rate, especially with appropriate preventive service, is very small, this factor is not considered in designing and is often used as a marketing argument. The rightness of this approach is justified by the industry information materials. Meanwhile, in our opinion, the dependability of ASRS should be treated much more broadly since the system is an expensive component of the warehouse facility and cannot operate separately. This category includes technical reliability of components, condition of cargo units, material assignment (slotting), spatial configuration of the rack system and handling devices, configuration of conveyor system, automation logics, and adaptive algorithms. When these factors are mixed into one with the structure of the material flow, then the system's dependability can be assessed.

For this article, the scope of ASRS solutions was limited to fully automated, combined systems of storage and internal transport consisting of stationary racks (single or double depth fixed-aisle system), stacker cranes equipped with a single or multi-seat fork carriage, a system of conveyors delivering and retrieving units from delivery and collection points, a system of sensors and identification devices, and possible connecting elements. Cranes use the single or combined transport cycles according to the adopted work logic. Carousels, vertical lift modules, and other forms of ASRS are excluded from this

study. A system defined in this way can be treated as a technical system characterized by certain reliability and dependability in the face of a logistic task.

The dependability of the automated storage and retrieval system should be considered concerning the following technological and organizational issues constituting the grounds for the problem formulation:

1. Availability of handling elements of ASRS (stacker cranes, conveyor systems, sensor systems, control systems).
2. Technical condition of ASRS devices and components (drives, control modules, construction frames, power supply).
3. Quality (condition) of handled logistic units (pallets, plastic containers, boxes) and its influence on handling processes.
4. Slotting patterns resulting from warehouse activity profiling and conditioning the flow congestion.
5. Configuration of structural components of ASRS (racking system, aisles, number of cranes, types of fork carriage, crane transfers).
6. Logic of operation.
7. Efficiency of low-level components of ASRS.
8. Information flow irregularities.
9. Material flow irregularities and accumulations resulting from orders structure.

Increased dependability of technical systems requires installed resources that potentially increase its cost or reduce its performance. So, the system's dependability can be influenced by its configuration and scale. Common methods for governing the dependability of the ASRS are as follows:

1. Technological redundancy:
 - Increasing the number of stacker cranes leads to an increased number of working aisles at the expense of the aisles' length and/or height.
 - Permanent assignment of stacker cranes to the aisles or using the transfer bridges and sliding mechanism to move the cranes between the corridors.
 - Use of multi-unit fork carriages.
 - Using single-deep racking systems instead of the double- or more deep lanes.
 - Universal and reconfigurable conveyor systems with redundant passages between main transport routes.
 - Doubled feeding system.
2. Material handling support systems:
 - The restrictive material carriers' quality policy (pallets, containers, boxes) when using units exchanged within the supply chain.
 - Advanced sensors systems detecting units bent out of shape or damaged.
 - Dedicated plastic containers or trays for material handling.
3. Slotting techniques:
 - Representing most popular or key SKUs in more than one aisle.
 - Functional division of the ASRS area into independent warehouse instances (two or more) in which all material groups (family groups) are independently represented.
 - Applying standard material assignment procedures based on warehouse activity profiling.
4. ASRS's place in the material flow organization:
 - Reduction of material flows pile up against the ASRS by rational work plan.
 - Equal load on individual working aisles (related to slotting).
 - Rationalization of the ASRS work schedule.

Redundancy always must be confronted with the effectiveness of the system. Typical ASRS solutions use one stacker crane in one aisle, so the number of stacker cranes equals the number of aisles. Such a configuration, with high relative technical reliability of devices, gives satisfactory results, simplifies the system, reduces the space require-

ment due to the lack of transfer mechanisms, and shortens the average operation time.

Multi-unit fork carriage enables task interleaving and increases system efficiency while maintaining partial efficiency of the stacker crane in non-critical damage to the handling device. The fork carriage is perceived to be quite vulnerable to damage, especially when interacting with a damaged load unit.

The use of single-deep racks ensures full stock selectivity in the ASRS area, which may be important in case of damage to the handling elements. It leads to a significant increase in the number of stacker cranes and space, but in case of failure of one of the devices, the cranes in other working aisles have access to the units of required material. Of course, the use of such a configuration requires an economic calculation of profitability. It is also strictly dependent on the number of SKUs and the number of material groups to be handled.

Conveyor systems are the second key component of ASRS supplying and receiving units from the ASRS. Conveyors can be configured in various ways. In most cases, the mainline system performs material flow, and the input and output separation is realized directly in front of the stacker cranes. To increase the reliability of the conveyor system, it is necessary to introduce the possibility of changing the flow direction of the selected conveyor sections (quite difficult to implement) and to place additional connections that will bypass damaged or congested places on the network. For warehouse process reasons, separated supply and receiving systems are used, as well as duplicated systems.

Practitioners report that potential failures in ASRS are often associated with poorly formed material units that lose stability, shape, or structural integrity during handling. This causes blocking of units in conveyor systems, stacker cranes and racks, damage to the installation, and requires operator intervention. Advanced sensor systems built into the conveyor network detect and withdraw damaged units to avoid problems, or manual quality control stations are used. Such systems increase the cost of installation but eliminate downtime caused by material quality problems. Another solution in this area is dedicated additional material carriers like a doubled pallet, plastic container, or tray, which are easily operable by the system but require additional handling and space.

The last of the essential techniques for increasing the dependability of ASRS is tailored slotting. In ASRS, apart from failures in power or control systems, single installation elements are damaged, making one of the working cranes inoperable. The other ones are functional. For this reason, it is important to represent all the key products in more than one place in the ASRS. Of course, solutions in this area must consider the number and type of products and warehouse activity profiles.

ASRS's place in the material flow organization may also impact the dependability of its work. The uneven workload of the system may temporarily exceed the efficiency of individual working aisles and conveyor systems supplying them. This, in turn, will cause congestion and, in the case of simplified control algorithms, may interfere with the operation of other ASRS elements. It is also important to maximize the available work time of the ASRS, which results from the schedule of the warehouse process.

4. ASRS dependability measures

The dependability of ASRS cannot be measured without the context of the warehouse system in which it works. Synthetic measures should be used to address the above-mentioned factors holistically and at the same time fit into the superior assessment of the warehouse process through OTIFEF (On-time, in-full, error-free) or POR (Perfect Order Rate). The OTIFEF measure is described in detail in [22] and usually if formulated separately for inbound and outbound processes since these processes have a little correlation in short time (daily regime). Still, it can be formulated as the probability of handling all periodical (daily) supplies and shipments on time and free of qualitative and quantitative errors or the percent of all supplies and

shipments handled in a standard way and in line with perfect-order requirements [23]. This measure can be reduced to the needs of the ASRS assessment to the time-related component since qualitative and quantitative errors are not generic to the automatic solutions.

The impact of ASRS operation on the OTIFEFF of the entire warehouse can be significant, especially when it feeds the material-to-human picking systems (wave picking) or direct shipments in the same-business day model. A delay in delivery of a single sku delays the execution of the entire order. In extreme cases, the order will be shipped incomplete if ASRS cannot deliver the material before the time window pass. This is strongly related to the warehouse process scheduling problem (as referred in [28]):

$$OTIFEFF = P_{OT} \cdot P_{IF} \cdot P_{EF} \quad (1)$$

where: P_{OT} , P_{IF} , and P_{EF} are the probability of handling all planned shipments on-time, in-full, and with no errors respectively.

To evaluate the P_{OT} component for ASRS the relation between resources R put into process realization and volume of orders must be found. Efficient resources assigned to ASRS will increase plausibility of immediate put-away and retrieval – dependability but cost more. Figure 1 shows the exemplary warehouse process in which ASRS is responsible for replenishing the picking area and directly outbound area with materials under the customer's orders. Distribution of resources constituting the dependability of ASRS will then influence the total order realization time t_3 :

$$t_3 = \max\{E(T_{RP},(R_{RP})), E(T_p,(R_p)), E(T_{RS},(R_{RS})), E(T_{SCP},(R_{SCP})), E(T_L,(R_L))\} \quad (2)$$

where:

- $E(T_{RP},(R_{RP}))$ – expected time of retrieving materials from ASRS with resources R_{RP} ,
- $E(T_p,(R_p))$ – expected time of picking in picking area with resources R_p ,
- $E(T_{RS},(R_{RS}))$ – expected time of retrieving from ASRS for direct shipment with resources R_{RS} ,
- $E(T_{SCP},(R_{SCP}))$ – expected time of sorting, consolidation and packing with resources R_{SCP} ,
- $E(T_L,(R_L))$ – expected time of loading materials with resources R_L ,

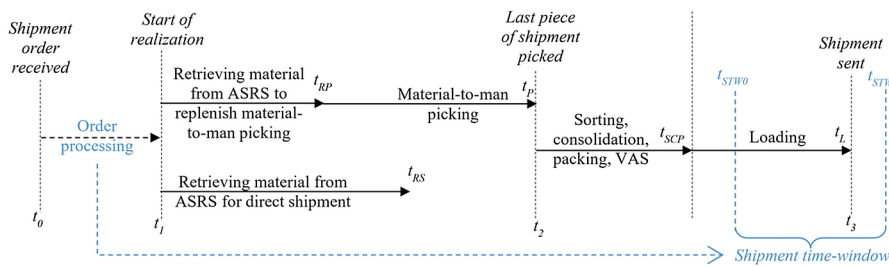


Fig. 1. Warehousing process using ASRS for order realization

In Figure 1, only the outbound processes are shown. In the analysed system, inbound processes requiring the same resources additionally load the ASRS. However, they are not directly responsible for the perfect-order-rate perceived by the client and then the dependability. Thus, inbound processes will affect the execution time of outbound processes, which will be considered in the simulation model.

Resources R reduced to their monetary value include ASRS equipment, mainly stacker cranes, which impact the productivity of the system. Resources influence directly handling potential (performance) and then the dependability of the system:

$$R = R_{RP} + R_P + R_{RS} + R_{SCP} + R_L$$

The above equation includes all resources in the analysed warehouse process. Still, if the resources not assigned to the ASRS are reduced to constant values, then it is possible to control the dependability of the warehousing system through the ASRS configuration. Then two tangled general criteria functions are used:

$$t_3 - t_1 \rightarrow \min \quad (3)$$

$$R \rightarrow \min \quad (4)$$

bounded by the constrain:

$$t_{STW0} \leq t_3 \leq t_{STW1} \quad (5)$$

where t_{STW0} and t_{STW1} are the start and the end moments of shipment time-window resulting from external to warehouse process conditions.

Therefore, the operation time is the main factor influencing the dependability of ASRS and, therefore, will be the basic factor tested in the simulation experiment.

5. Assumptions for the simulation experiment

The experiments were carried out in the simulation model prepared in *FlexSim* – 3D simulation modeling and analysis software (v. 21.2.0). Prepared model allows for simulation of single-deep ASRS of any configuration and with any workload.

Model uses 1 to 10 work aisles with fixed or transferred cranes, single-deep racking system, two in/out conveyor systems for separated or combined delivery and retrieval, MTBF and MTTR functions for all elements and range of slotting patterns (Figures 2 and 3).

The configuration of the experimental system is based on:

- 20 single-deep rack walls (20 bays, 12 levels, 3 slots per rack cell) for 1200x800 EUR1 pallet units with a maximum height of 1200 mm,
- 1 to 10 pallet cranes ($V^{\max} = 1,6 \text{ m/s}$, acceleration / deceleration $A = 0,3 \text{ m/s}^2$),
- 1 transfer for cranes ($V^{\max} = 1 \text{ m/s}$, acceleration / deceleration $A = 0,2 \text{ m/s}^2$),

- upper conveyor system (only for optional separated collection, $V^{\max} = 1 \text{ m/s}$),
- bottom conveyor system (collection and delivery, $V^{\max} = 1 \text{ m/s}$).

Stacker cranes are assigned to working aisles, but the activated stacker cranes are less than 10, the transfer moves them between the working aisles, searching for the nearest stacker crane at idle. The system of conveyors delivering and collecting units from racks is either integrated or separated.

The conveyor system allows the circulation of units addressed into the racks. If it is not possible for the unit to enter the conveyor segment

supplying a given rack, the unit will perform a maximum of 3 loops, and after the third attempt, it will leave the system unhandled.

The examined ASRS supplies the dynamic order picking system with required materials and deposits the units leaving this system. It is also used to buffer homogeneous units directly from delivery and releases units outgoing directly to customers. Therefore, delays in the put-away or retrieval of units by ASRS will impact the remaining components of the warehouse process and thus on OTIFEFF.

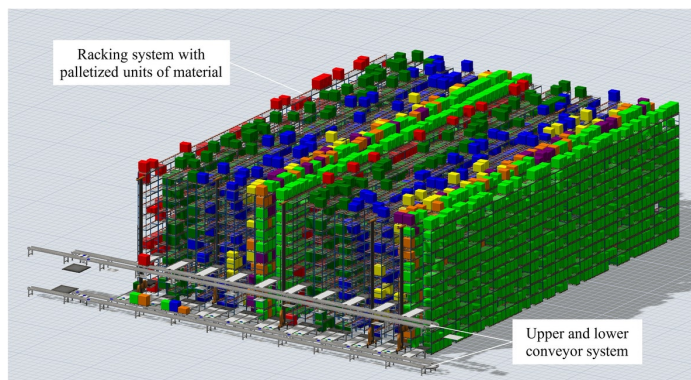


Fig. 2. General view of the ASRS model in operation

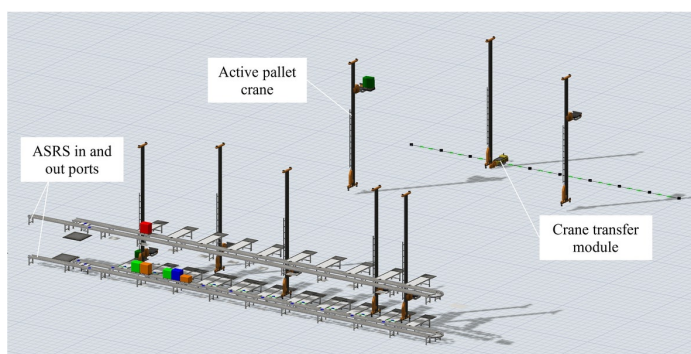


Fig. 3. General view of the pallet cranes in operation

According to the ABC principle, the material in the system was divided into 7 material groups with different turnover and initial stock. The system operates 16 hours a day, while the schedules for deliveries retrievals assume an uneven flow at selected hours (Table 1). It was assumed that the system realizes on average 300 orders of 6 pallets (SKUs) each. Initial stock represents the material structure in line with the distribution of material groups and their parameters (Table 2).

Table 1. Material flow schedule

| Hours | % of daily delivery | % of daily retrieval |
|---------------------------|---------------------|----------------------|
| 8.00 – 9.00 ^{*)} | 5 | 1 |
| 9.00 – 10.00 | 5 | 7 |
| 10.00 – 11.00 | 5 | 7 |
| 11.00 – 12.00 | 10 | 7 |
| 12.00 – 13.00 | 20 | 7 |
| 13.00 – 14.00 | 15 | 7 |
| 14.00 – 15.00 | 10 | 10 |
| 15.00 – 16.00 | 10 | 10 |
| 16.00 – 17.00 | 1 | 10 |
| 17.00 – 18.00 | 5 | 8 |
| 18.00 – 19.00 | 5 | 8 |
| 19.00 – 20.00 | 3 | 8 |
| 20.00 – 21.00 | 2 | 5 |
| 21.00 – 22.00 | 2 | 5 |
| 22.00 – 23.00 | 1 | 0 |
| 23.00 – 24.00 | 1 | 0 |

^{*)} Intervals are rounded to whole hours.

Table 2. Simulation scenarios

| Group of material | % of stock | % of flow (% of total number of units) | Number of SKUs in the group | Initial stock [units] |
|-------------------|------------|--|-----------------------------|-----------------------|
| A | 1 | 10 | 10 | 239 |
| B | 4 | 25 | 40 | 998 |
| C | 10 | 30 | 100 | 813 |
| D | 10 | 10 | 100 | 389 |
| E | 10 | 7 | 100 | 595 |
| F | 10 | 7 | 100 | 345 |
| G | 55 | 11 | 550 | 1921 |

All elements of equipment are described by reliability functions: Mean time between failures (MTBF) and Mean time to repair (MTTR) as it results from [24] (Table 3).

To illustrate the aspects of ASRS dependability discussed above, the 160 simulation runs for 40 scenarios were done. The spectrum of scenarios is based on a changing number of active cranes (2, 5, 7, and 10, respectively), the use of a separate entry and exit system, and five variants of product slotting patterns (Figure 4).

1. Random location (SP1).
2. Volume-based product location along work aisles (SP2).
3. Volume-based left-to-right product location (SP3).
4. Two separated storage areas with a random location (SP4).
5. Two separated storage areas with volume-based left-to-right product location (SP5).

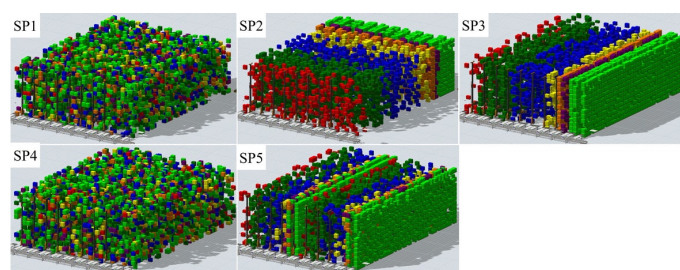


Fig. 4. Stock visualization for different slotting patterns (SP)

Selected slotting patterns will reveal the bottlenecks of the system affecting its actual dependability.

6. ASRS dependability simulation

The simulation was presented in a one-day and monthly regime to show the impact of potential damage to the operating components on the system's dependability. During the simulation, the basic parameters determining ASRS usability in the warehouse process were examined: the average put-away time (Table 4 and Figure 5) and the

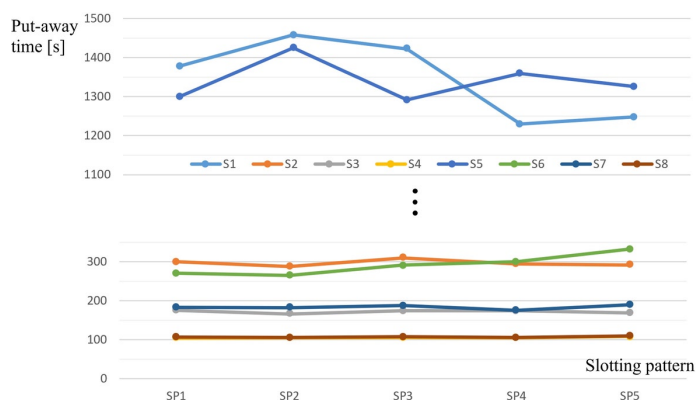


Fig. 5. Mean (95% confidence interval) of average put-away time

Table 3. Statistical distributions of MTBF and MTTR

| Type of equipment | Down time [s] | Up time [s] ^{*)} |
|-----------------------|----------------------------------|--|
| Cranes | Uniform (between 1800 and 28800) | Exponential (location 14400, scale 5400000) |
| Cranes control system | Uniform (between 900 and 57600.) | Exponential (location 403202, scale 1612800) |
| Transfer | Uniform (between 1800 and 5400) | Exponential (location 52200, scale 1607400) |
| Lower conveyor set | Uniform (between 1800 and 5400) | Exponential (location 52200, scale 1607400) |
| Upper conveyor set | Uniform (between 1800 and 5400) | Exponential (location 52200, scale 1607400) |

^{*)} First failure time equal to up time distribution.

Table 4. Mean (95% confidence interval) of average put-away time [s]

| Scenario | No of cranes | Upper conveyors | SP1 | SP2 | SP3 | SP4 | SP5 |
|----------|--------------|-----------------|------------------|------------------|------------------|------------------|------------------|
| S1 | 2 | Not used | 1378,16 ± 105,74 | 1458,42 ± 210,63 | 1422,21 ± 314,82 | 1229,70 ± 206,57 | 1248,46 ± 161,95 |
| S2 | 5 | Not used | 299,77 ± 63,37 | 287,63 ± 24,97 | 309,93 ± 81,66 | 294,22 ± 40,43 | 291,53 ± 40,65 |
| S3 | 7 | Not used | 175,16 ± 25,62 | 166,30 ± 16,59 | 174,81 ± 14,64 | 174,17 ± 16,31 | 168,62 ± 11,72 |
| S4 | 10 | Not used | 104,31 ± 3,50 | 104,70 ± 6,27 | 105,67 ± 2,73 | 104,76 ± 2,69 | 108,34 ± 2,80 |
| S5 | 2 | Used | 1300,00 ± 248,10 | 1424,91 ± 538,32 | 1291,54 ± 166,68 | 1359,02 ± 123,95 | 1325,65 ± 146,61 |
| S6 | 5 | Used | 270,50 ± 22,33 | 264,58 ± 17,18 | 291,12 ± 46,92 | 300,37 ± 85,13 | 332,45 ± 71,26 |
| S7 | 7 | Used | 183,13 ± 28,61 | 182,47 ± 19,90 | 188,11 ± 12,80 | 175,65 ± 22,92 | 189,79 ± 7,01 |
| S8 | 10 | Used | 106,99 ± 3,49 | 106,28 ± 4,86 | 107,45 ± 2,88 | 106,42 ± 2,57 | 110,13 ± 3,79 |

Table 5. Mean (95% confidence interval) of average retrieval time [s]

| Scenario | No of cranes | Upper conveyors | SP1 | SP2 | SP3 | SP4 | SP5 |
|----------|--------------|-----------------|-------------------|--------------------|-------------------|-------------------|-------------------|
| S1 | 2 | Not used | 3753,93 ± 1714,11 | 2759,37 ± 1555,10 | 3357,70 ± 1502,94 | 6315,52 ± 5374,42 | 4713,04 ± 2392,40 |
| S2 | 5 | Not used | 1214,31 ± 764,99 | 1261,35 ± 689,56 | 1102,20 ± 884,82 | 1259,55 ± 711,66 | 1241,70 ± 770,32 |
| S3 | 7 | Not used | 333,29 ± 43,13 | 305,87 ± 89,77 | 373,19 ± 19,95 | 352,29 ± 70,50 | 303,33 ± 92,11 |
| S4 | 10 | Not used | 174,12 ± 2,41 | 172,93 ± 3,37 | 177,72 ± 5,31 | 176,09 ± 3,39 | 175,61 ± 2,86 |
| S5 | 2 | Used | 8512,85 ± 5651,78 | 11458,12 ± 7652,40 | 8675,32 ± 4258,15 | 7279,63 ± 3787,02 | 8321,44 ± 4176,63 |
| S6 | 5 | Used | 812,99 ± 503,59 | 711,26 ± 559,23 | 1094,83 ± 1042,79 | 1091,60 ± 945,93 | 1529,83 ± 596,95 |
| S7 | 7 | Used | 336,96 ± 52,08 | 357,69 ± 54,87 | 354,88 ± 69,20 | 325,58 ± 118,36 | 337,84 ± 73,45 |
| S8 | 10 | Used | 174,90 ± 2,91 | 173,29 ± 3,20 | 178,49 ± 5,24 | 176,93 ± 3,06 | 176,57 ± 2,80 |

average retrieval time (Table 5 and Figure 6), the number of units handled in a given time, and the number of delayed units (Table 6).

The above data present dependencies between ASRS configuration (assigned resources R) and slotting rules at constant loads and device reliability functions. The most important measure for the ASRS dependability is the retrieval and depositing time (Figures 5 and 6). These parameters determine the time component of the material release process, which is crucial for the Perfect Order Rate index, and thus for the quality of customer service (conf. [21]).

Following the assumptions given in point X, the unit service time in the ASRS depends on the speed of unit movement, the availability of the cranes in the working corridor, congestion in the elements of the conveyor system, and technical reliability of the system components.

As shown in Figure 5, the average put-away time is strictly dependent on the number of stacker cranes in the system. When 2 of 10 cranes (scenarios S1 and S5) are used, the unit put-away times range from appr. 1 200 s (20 min) to appr. 1 450 s (25 min), which results from the lack of available stacker crane in the corridor and the need

to move it between corridors. This causes the congestion of units in the conveyor system, which pushes out units from the system after 3 unsuccessful attempts (Table 6).

By increasing the number of stacker cranes to 5, 7, and 10 respectively, the access time is reduced. For 5 of 10 stacker cranes, the congestion in the conveyor system is not visible.

The separation of the input and output conveyors (scenarios S5 to S10) reduces the put-away time with a small number of stacker cranes but does not significantly affect this time with 5 or more cranes.

Average retrieval time is shaped by the same principles (Figure 6). The very long retrieval time is particularly exposed in scenarios with 2 of 10 stacker cranes (S1 and S5), which is an extreme case reached 3,2 hours with a common conveyor system for entry and exit. This is an obvious aberration resulting from the extreme congestion of units, which makes it impossible to complete the ASRS logistics task. As the number of stacker cranes increases, times are normalized. Longer times of retrieval operations result indirectly from the logic of the

Table 6. Mean (95% conf. int.) of average number of put-away / retrieved / not served units

| Scenario | No of cranes | Upper conveyors | SP1 | SP2 | SP3 | SP4 | SP5 |
|----------|--------------|-----------------|-------------------------|------------------------|-------------------------|------------------------|-------------------------|
| S1 | 2 | Not used | 440.25/ 416.50/ 273.50 | 357.50/ 332.75/ 178.25 | 398.75/ 380.50/ 229.00 | 835.75/ 570.50/ 658.00 | 513.50/ 500.50/ 388.75 |
| S2 | 5 | Not used | 1787.25/ 1028.00/ 24.50 | 1791.50/ 998.00/ 20.25 | 1779.25/ 1120.50/ 32.50 | 1791.00/ 990.25/ 20.75 | 1782.00/ 962.75/ 29.75 |
| S3 | 7 | Not used | 1810.00/ 1305.00/ 1.75 | 1808.25/ 1019.25/ 3.50 | 1809.00/ 1346.50/ 2.75 | 1809.00/ 1540.00/ 2.75 | 1808.25/ 1149.75/ 3.50 |
| S4 | 10 | Not used | 1811.75/ 1795.25/ 0.00 | 1811.75/ 1795.25/ 0.00 | 1811.75/ 1795.25/ 0.00 | 1811.75/ 1795.25/ 0.00 | 1811.75/ 1795.25/ 0.00 |
| S5 | 2 | Used | 705.50/ 697.50/ 569.00 | 645.25/ 986.75/ 511.00 | 608.50/ 714.25/ 454.00 | 501.00/ 628.00/ 365.25 | 546.00/ 675.75/ 435.75 |
| S6 | 5 | Used | 1794.75/ 727.75/ 17.00 | 1801.25/ 655.25/ 10.50 | 1786.00/ 845.75/ 25.75 | 1778.00/ 891.00/ 33.75 | 1766.25/ 1243.50/ 45.50 |
| S7 | 7 | Used | 1809.00/ 1537.50/ 2.75 | 1809.25/ 1440.50/ 2.50 | 1807.50/ 1795.00/ 4.25 | 1810.25/ 1249.00/ 1.50 | 1807.75/ 1791.25/ 4.00 |
| S8 | 10 | Used | 1811.75/ 1795.25/ 0.00 | 1811.75/ 1795.25/ 0.00 | 1811.75/ 1795.25/ 0.00 | 1811.50/ 1795.25/ 0.25 | 1811.75/ 1795.25/ 0.00 |

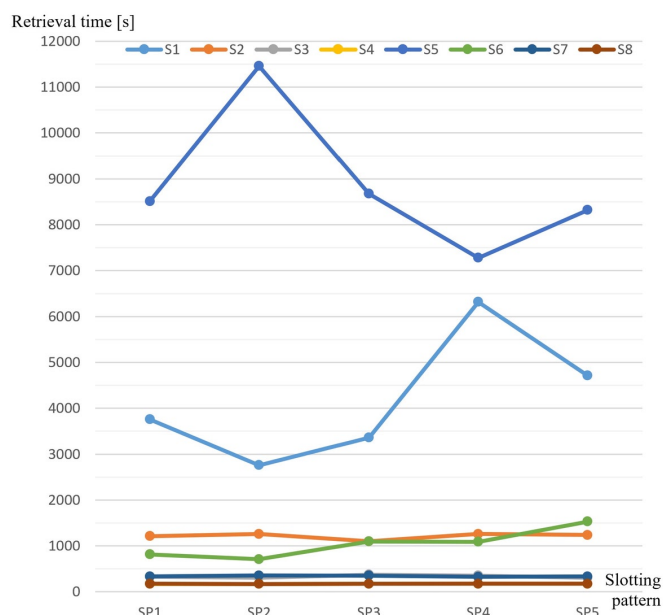


Fig. 6. Mean (95% confidence interval) of average retrieval time

stacker crane. The crane to be moved between the corridors will perform an average of 2 put-away operations per 1 retrieval operation.

Slotting scenarios based on the random distribution of the assortment (SP1) in the locations are characterized by the shortest operation times, which results from the logic of the stacker crane operation. The crane changing the corridors performs combined cycles, and traverses the entire corridor length regardless of the picking address. Slotting patterns using volume-based material assignment (especially SP2) allow for a slight reduction in the operation time, but it is related to the logic of the cranes.

According to the literature on the subject, the technical reliability of the ASRS elements (stacker cranes, transfer, control system, conveyors) does not have a noticeable effect on the ASRS opera-

tion. It is clear especially when scheduled maintenance programs are executed outside the regular work time. Recorded occurrences of damage and recovery times did not affect the reliability of the entire ASRS in this case.

7. Conclusions

The article presents a discussion on reliability in logistic systems, which cannot always be used as a measure for the assessment of warehouse technologies. Complex storage systems, especially multi-unit integrated automatic solutions such as Automated Storage and Retrieval Systems, pose new challenges in measuring their reliability. While it is quite clear on the technical level, the complex conditions of the surrounding logistics process make the assessment of ASRS solutions difficult. A much better solution turns out to be the use of dependability measures, which also consider non-structural factors of warehouse technology, especially related to work patterns and allocated labour resources (cost-effectiveness).

The ASRS configuration and allocated resources affect its performance, especially at high workloads. They must be considered as important factors forming the dependability of ASRS and the entire warehouse process.

The simulation studies showed the influence of configuration factors and organizational factors such as material slotting on expected retrieval and put-away times, which in turn are of great importance for the perfect-order-rate of the entire warehousing process.

Therefore, the approach used in practice presented in the Introduction section seems to be right. In this approach, the reliability measures are abandoned in warehouse automation in favour of efficiency measures. However, in this case, they should also be related to specific working conditions, which is postulated in this article.

Further research in this area should include developing a catalogue of standard factors (and their measures) influencing the dependability of ASRS as components of a warehouse system focused on the execution of customer orders.

References

1. Atmaca E, Ozturk A. Defining order picking policy: A storage assignment model and a simulated annealing solution in AS/RS systems. *Applied Mathematical Modelling* 2013; 37: 5069–5079, <http://dx.doi.org/10.1016/j.apm.2012.09.057>.
2. Azadeh K, De Koster R, Roy D. Robotized and automated warehouse systems: review and recent developments *Transportation Science* 2019; 53(4): 917–945, <https://doi.org/10.1287/trsc.2018.0873>.

3. Baghalian A, Rezapour S, Farahani RZ. Robust supply chain network design with service level against disruptions and demand uncertainties: A real-life case. *European Journal of Operational Research* 2013; 227(1): 199–215, 10.1016/j.ejor.2012.12.017.
4. Barnes E, Dai J, Deng S, Down D, Goh M, Lau H C, Sharafali M. On the Strategy of Shupply Hubs for Cost Reduction and Responsiveness. White Paper, The Logistics Institute – Asia Pacific 2003. National University of Singapore.
5. Boysen N, Stephan K. A survey on single crane scheduling in automated storage/retrieval systems. *European Journal of Operational Research* 2016; 254: 691–704, <https://doi.org/10.1016/j.ejor.2016.04.008>.
6. Brezovnik S, Gotlih J, Balić J, Gotlih K, Brezočnik M. Optimization of an Automated Storage and Retrieval Systems by Swarm Intelligence. *Procedia Engineering* 2015; 100: 1309 – 1318.
7. Bukowski L, Feliks J. A unified model of systems dependability and process continuity for complex supply chains. In: Nowakowski T. et al. (eds) *Safety and Reliability: Methodology and Applications*, CRC Press Taylor & Francis Group, 2015: 2395-2403.
8. Bukowski L. System of systems dependability – Theoretical models and applications examples. *Reliability Engineering & System Safety* 2016; 151: 76–92.
9. Chung S H, Chan H K, Chan F T S. A modified genetic algorithm for maximizing handling reliability and recyclability of distribution centers. *Expert Systems with Applications* 2013; 40(18): 7588–7595.
10. Ekren B Y, Heragu S S. Simulation based performance analysis of an autonomous vehicle storage and retrieval system, *Simulation Modelling Practice and Theory* 2011; 19: 1640–1650.
11. Ekren B Y, Sari Z, Lerher T. Warehouse Design under Class-Based Storage Policy of Shuttle-Based Storage and Retrieval System. *IFAC-PapersOnLine* 2015; 48(3): 1152-1154.
12. Fazlollahtabar H, Saidi-Mehrabad M. Optimising a multi-objective reliability assessment in multiple AGV manufacturing system. *International Journal of Services and Operations Management* 2013; 16(3): 352–372.
13. Gademann A N. Optimal routing in an automated storage/retrieval system with dedicated storage. *IIE Transactions* 1999; 31(5): 407–415.
14. Gagliardi J-P, Renaud J, Ruiz A. Models for automated storage and retrieval systems: a literature review. *International Journal of Production Research* 2012; 50(24): 7110-7125, <http://dx.doi.org/10.1080/00207543.2011.633234>.
15. Hachemi K, Besombes B. Integration of products expiry dates in optimal scheduling of storage/retrieval operations for a flow-rack AS/RS. *International Journal of Industrial and Systems Engineering* 2013; 15(2): 216–233, 10.1504/IJISE.2013.056097.
16. Haj Shirmohammadi A. *Programming maintenance and repair. Technical management in industry* 2002, 8th edition. Esfahan: Ghazal Publishers.
17. International Electrotechnical Commission, Electropedia. 192 Dependability, IEC ref 192-01-22. Available online: <https://www.electropedia.org> (accessed on Jul 27, 2021).
18. Jachimowski R, Gołębiowski P, Izdebski M, Pyza D, Szczepański E. Designing and efficiency of database for simulation of processes in systems. Case study for the simulation of warehouse processes. *Archives of Transport* 2017; 41(1): 31-42, 10.5604/01.3001.0009.7380.
19. Jacyna M, Semenov I. Models of vehicle service system supply under information uncertainty. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2020; 22 (4): 694–704, <http://dx.doi.org/10.17531/ein.2020.4.13>.
20. Jacyna M, Wasiak M, Bobiński A. SIMMAG3D as a tool for designing of storage facilities in 3D. *Archives of Transport* 2017; 42(2): 25-38, 10.5604/01.3001.0010.0525.
21. Jacyna-Golda I, Kłodawski M, Lewczuk K, Łajszczak M, Chojnacki T, Siedlecka-Wójcikowska T. Elements of perfect order rate research in logistics chains. *Archives of Transport* 2019; 49(1): 25-35, 10.5604/01.3001.0013.2771.
22. Jacyna-Golda I, Lewczuk K, The method of estimating dependability of supply chain elements on the base of technical and organizational redundancy of process. *Eksplotacja i Niezawodność – Maintenance and Reliability* 2017; 19 (3): 382–392, <http://dx.doi.org/10.17531/ein.2017.3.9>.
23. Kłodawski M, Lewczuk K, Jacyna-Golda I, Żak J. Decision making strategies for warehouse operations. *Archives of Transport* 2017; 42(1): 43–53, 10.5604/01.3001.0009.7384.
24. Kulwiec R, Ray Kulwiec Associates. Reliability of automated storage and retrieval systems (AS/RS) a White Paper. 2007. Available online: http://www.pnkreis.com/images/column_1448348195/asrswitepaper3%20mhi.pdf (accessed on 30 Jun 2021).
25. Lerher T, Ekren Y B, Sari Z, Rosi B. Simulation analysis of shuttle based storage and retrieval systems. *International Journal of Simulation Modelling* 2015; 14(1): 48-59, 10.2507/IJSIMM14(1)5.281.
26. Lerher T, Ficko M, Palčić I, Throughput performance analysis of Automated Vehicle Storage and Retrieval Systems with multiple-tier shuttle vehicles. *Applied Mathematical Modelling* 2021; 91: 1004–1022, <https://doi.org/10.1016/j.apm.2020.10.032>.
27. Lewczuk K, Dependability issues in designing warehouse facilities and their functional areas. *Journal of KONBiN* 2016; 2(38): 201–228, 10.1515/jok-2016-0024.
28. Lewczuk K. The concept of genetic programming in organizing internal transport processes. *Archives of transport* 2015; 34(2): 61-74, 10.5604/08669546.1169213.
29. Liu T, Gong Y, De Koster R B M. Travel time models for split-platform automated storage and retrieval systems. *International Journal of Production Economics* 2018; 197: 197–214, <https://doi.org/10.1016/j.ijpe.2017.12.021>.
30. Manzini R, Gamberi M, Regattieri A. Design and control of an AS/RS. *The International Journal of Advanced Manufacturing Technology* 2006; 28: 766–774, 10.1007/s00170-004-2427-6.
31. Marchet G, Melacini M, Perotti S, Tappia E. Analytical model to estimate performances of autonomous vehicle storage and retrieval systems for product totes. *International Journal of Production Research* 2012; 50(24): 7134–7148, <https://doi.org/10.1080/00207543.2011.639815>.
32. Marchet G, Melacini M, Perotti S, Tappia E. Development of a framework for the design of autonomous vehicle storage and retrieval systems. *International Journal of Production Research* 2013; 51(14): 4365–4387, 10.1080/00207543.2013.778430.
33. Nativ D J, Cataldo A, Scattolini R, De Schutter B. Model Predictive Control of an Automated Storage/Retrieval System. *IFAC-PapersOnLine* 2016; 49-12: 1335–1340, 10.1016/j.ifacol.2016.07.745
34. Neo H Y, Xie M, Tsui K L. Service quality analysis: case study of a 3PL company. *International Journal of Logistics Systems and Management* 2004; 1(1): 64–80.
35. Ning Z, Lei L, Saipeng Z, Lodewijks G, An efficient simulation model for rack design in multi-elevator shuttle-based storage and retrieval system. *Simulation Modelling Practice and Theory* 2016; 67: 100–116, <http://dx.doi.org/10.1016/j.simpat.2016.03.007>.
36. Nowakowski T. Analysis of possibilities of logistics systems reliability assessment. *Safety and Reliability for managing risk* 2006; 3. Leiden:

Taylor and Francis.

37. Nowakowski T. Models of uncertainty of operation and maintenance information. *Zagadnienia Eksploatacji Maszyn* 2000; 35(2): 143–150.
38. Nowakowski T. Reliability model of combined transportation system. *Probabilistic Safety Assessment and Management*. Spitzer C, Schmocker U, Dang V N (ed.). London: Springer, 2004.
39. Quigley J, Walls L. Trading reliability targets within a supply chain using Shapley's value. *Reliability Engineering & System Safety* 2007; 92(10): 1448–1457.
40. Ripple J. Automated Storage and Retrieval Systems: The Impact of Load Condition on ASRS Reliability and Availability 2019. Available online at <https://www.linkedin.com/pulse/automated-storage-retrieval-systems-impact-load-condition-john-ripple/> (accessed on 30 Jun 2021).
41. Roodbergen K J, Vis I F A. A survey of literature on automated storage and retrieval systems. *European Journal of Operational Research* 2009; 194(2): 343–362, 10.1016/j.ejor.2008.01.038.
42. Sarker B R, Babu P S. Travel time models in automated storage/retrieval systems: A critical review. *International Journal of Production Economics* 1995; 40: 173–184.
43. Sohn S Y, Choi I S, Fuzzy QFD for supply chain management with reliability consideration. *Reliability Engineering & System Safety* 2001; 72(3): 327–334.
44. Szaciłło L, Jacyna M, Szczepański E, Izdebski M. Risk assessment for rail freight transport operations. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2021; 23 (3): 476–488, <http://doi.org/10.17531/ein.2021.3.8>.
45. Vasili M, Hong T S, Homayouni S M, Ismail N. A statistical model for expected cycle time of SP-AS/RS: an application of Monte Carlo simulation. *Applied Artificial Intelligence* 2008; 22 (7–8): 824–840, <https://doi.org/10.1080/08839510802374841>.
46. Werbińska-Wojciechowska S. The availability model of logistic support system with time redundancy. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2007; 3(35): 23–29.
47. Werbińska-Wojciechowska S. Time resource problem in logistics systems dependability modelling. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2013; 15(4): 427–433.
48. Lee Y K, Yang H S. A study for secure the reliability of automated warehouse system, *Journal of Digital Convergence* 2016; 14 (10): 253–259, <https://doi.org/10.14400/JDC.2016.14.10.253>.
49. Yan R, Dunnett S J, Jackson L M. Reliability Modelling of Automated Guided Vehicles by the Use of Failure Modes Effects and Criticality Analysis, and Fault Tree Analysis. 5th Student Conference on Operational Research (SCOR'16) 2016. Hardy B, Qazi A, Ravizza S (eds); Article No. 2: 2:1–2:11, 10.4230/OASICS.SCOR.2016.2.
50. Yang D, Wu Y, Ma W. Optimization of storage location assignment in automated warehouse. *Microprocessors and Microsystems* 2021; 80: 103356, <https://doi.org/10.1016/j.micpro.2020.103356>.
51. Zhang Z, Liu Q, Lv C, Zhang L, Li S. An NSABC algorithm for multi-aisle AS/RS scheduling optimization Xiaohui Yan. *Computers & Industrial Engineering* 2021; 156: 107254, <https://doi.org/10.1016/j.cie.2021.107254>.

Article citation info:

Vaičiūnas G, Steišūnas S, Bureika G. Specification of estimation of a passenger car ride smoothness under various exploitation conditions. *Eksploracja i Niezawodność – Maintenance and Reliability* 2021; 23 (4): 719–725, <http://doi.org/10.17531/ein.2021.4.14>.

Specification of estimation of a passenger car ride smoothness under various exploitation conditions

Indexed by:



Gediminas Vaičiūnas^{a*}, Stasys Steišūnas^a, Gintautas Bureika^a

^aVilnius Gediminas Technical University, Faculty of Transport Engineering, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania


Highlights

- Sperling's comfort index dependence on the stiffness of passenger car suspension.
- Wheel flat impact on rail vehicle running gear vibration character.
- Guidance & stability of running gear with independently rotating wheels.
- Processing of the carbody acceleration amplitudes by Fourier transform method.

Abstract

The stability and smoothness of rolling stock running could be defined accurately by universal Sperling's comfort index. The divergences of variation of Sperling's comfort index of a passenger car under specific operating conditions of running gear are examining in this paper. Numerical simulations of a passenger car running with independently rotating wheels under various conditions have been performing. Gained results showed that divergences of the Sperling's comfort index variation are particularly significant due to running gear component oscillations in the horizontal plane (lateral direction). A field experiment of a passenger car with a solid (traditional) wheelset with a flat running surface proved this hypothesis. The obtained results of this experiment confirmed this assumption. Therefore, the study of the regularities of lateral oscillations of a passenger car is the logical direction of further research.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

railway transport, passenger car, running gear, independently rotating wheels, Sperling's comfort index, divergences, numerical simulation, software package UM.

1. Introduction

The improvement of rolling stock raises several problems of the mechanical wheel-rail interaction: the risk of derailment, the intensity of rolling surface wear and the discomfort caused to passengers by vibrations [12, 14, 30]. In typical cases, technical solutions are under development to eliminate these problems. However, there are also unsolved aspects of the problems as mentioned above.

The wheel conicity ensures the stability and guidance of rolling stock with solid (ordinary) wheelsets and uniform rotational speed, respectively higher or lower linear speed in contact with rails [23]. It does not work on rail vehicles with independently rotating wheels, so other methods are needed. The damping of rolling contact of wheel and rail in dry friction provided by the primary suspension dampers of cargo rolling-stocks is considered in numerical simulations performed by Polish scientist Piotrowski [22]. Noticeable that the stability and smoothness of rolling stock running influenced the wear intensity of running gear and track components [11, 26].

The proposed power-steering railway bogie consists of independently rotating wheels (IRWs) with a power-steering device. It enables us to eliminate steering vibration while realising ideal steering with slight power assist on curving [3, 18]. There is a proposed use of IRWs with inverse tread conicity to get self-steering ability without any complex bogie structure. The testing and numerical simulation results show that the proposed IRWs with inverse tread conicity

have good performance [27, 28]. The benefits of implementing active steering systems in railway vehicles mounting bogies with IRWs and outlines a design methodology for such systems are presented [21].

Noticeably that the parameters and characteristics of wheelsets with IRWs are regulated by law. In research, scientists also examine them, for example, the standard ISO 2631, EN 12299:2009 [10]. However, legal issues are not the subject of this research.

Solving rolling stock stability issues leads to passenger comfort issues, and peculiarities also occur here. One of them is passenger comfort in terms of vibrations. The Sperling's comfort index (SCI) is commonly used in scientific research to assess the passenger car ride smoothness in terms of vibrations. One of the main directions of railway development is to increase the running speed of trains. Naturally, research is usually carried out at high speeds, and the SCI is examined at high speeds (more than 160 km/h). However, with the development of rail transport, specific cases always occur, such as running vehicles with IRWs on small radius curves (less than 300 m radii). This refers to railway track repair works, where vehicles need to move from one track to another or manoeuvring in railway stations or tunnels. In this case, the speed is lower (there may be restrictions of 50 km/h and less). Passenger car ride smoothness is also essential here, and a study of the SCI for such issues is needed. The rolling surface of the wheels could be damaged when vehicles are running on poor quality railway tracks. With larger than the allowable damage, continued running on

(*) Corresponding author.

E-mail addresses: G. Vaičiūnas - gediminas.vaiciunas@vilniustech.lt, S. Steišūnas - stasys.steissunas@vilniustech.lt, G. Bureika - gintautas.bureika@vgtu.lt

rail vehicles is prohibited. However, the permissible extent of damage caused by vibrations affects passenger car ride smoothness and must be estimated by SCI.

Examples of the study of the impact of rolling stock wheels with damage on the rails and on the rail vehicle ride smoothness are described in scientific papers [2, 20]. The most common damage to the wheelset is the unevenness of the rolling surface, the wear of the flange, wheel flats and cracks. The unevenness of the wheel rolling surface of the wheels can be divided into three types according to their effect on the rail:

1. Unevenness causing impact and loss of contact (flats, bends, abrasions, cracks, etc.). The unevenness of the wheel rolling surface is usually characterised by the depth and the length of the flat.
2. Insulated irregularities increase the vertical impact of the wheel on the rail without loss of contact (uneven wear, "out-of-roundness", etc.).
3. Wheel flange damage. The flange prevents the wheelset from the derailment. A wheel is considered unusable and unsafe when its flange is critically thinned (equal to or less than 25 mm).

Problematic of wheel rolling surface are considered in the most publications about the long-term interaction between rolling stock running gear and track [8, 17, 4], the intensity of wheelset wear is also examined [1, 7, 13, 23]. The phenomena of wheelset wear have been extensively studied [1, 6, 15]. The wear of the wheel rolling surface is divided into even and uneven. Even wear is wear of the wheel rolling surface when the wheel rolling surface wears evenly (regular "circle"). Uneven wear of the wheel rolling surface differs from even wear in that the rolling surface wears unevenly ("out-of-roundness"), which increases the dynamic impact of the wheel on the rail [15, 25]. It is difficult to find such damage without removing the wheelset during a wagon inspection, as uneven wear can account for one-fifth or more of the total wheel surface [28].

The flats are the most common wheel running surface damages due to wheelset slip or jammed brake pads [5, 27, 31]. Flats result from wheel skidding, wheel jamming, or brake failure (especially during the wagon sorting on hubs). Flats occur in winter much more often than in summer. Mathematical models of the impact effect of wheelset with a flat on the rail have been discussed in the works of various scientists [5, 26].

The combination of short-term dynamics and long-term wear processes is a very complicated and unexplored phenomenon, but the influence of physical factors such as surface unevenness, material properties, or micro-crack intensity must be considered [24, 32]. In most scientific research, wear processes are usually simplified and conditioned only by frictional forces, and the dependence on plastic deformation and other processes influencing the formation of cracks are not considered [20, 16]. The study of wheelset damage observed that the damage formation process is a complicated and complex process. Finally, the analysis of wheelset damage shows that the safe and smooth movement of rolling stock is greatly influenced by the shape and condition of the rolling surface of the wheelset wheel [9].

Some research has been performed by scientists of Korea Railroad Institute to correlate various evaluation methods by using different vibration models [12]. The ride comfort indexes defined in ISO 2631 and EN 12299:2009 are commonly adopted in favour of the SCI method is seldom applied and discussed. Ride comfort in railway vehicles on a track with vertical irregularities was evaluated by implementing two different comfort indexes, corresponding to the EN 12299:2009 and SCI method, respectively [24]. The ride comfort level of passengers in two positions, sitting and standing, was compared using the EN 12299:2009 and SCI methods [19]. The Ride Comfort Index discussed in both studies is the Mean Comfort Index. Another frequently used Ride Comfort Index in EN 12299:2009 is called the Continuous Comfort Index. This index uses a quadratic

average (r.m.s) of the frequency weighted accelerations measured to evaluate the Mean Comfort [14]. Since the mean comfort is determined in the longitudinal, lateral, and vertical directions, respectively, and it has similarities to Sperling's comfort index.

Based on a comparative analysis of the methods in the literature, the SCI was selected by Authors as the most appropriate indicator to assess the running comfort of a passenger car. This study aims to provide different ways of SCI identification under specific operating conditions of passenger wagon, such as running on a small radius curve of a track or when the wheel running surface is damaged. In order to reduce the intensity of wear of the rolling stock wheel flange due to the friction on the track curves, the possibility of installing IRWs on the rolling stock (instead of the usual solid wheelsets) is investigated. Various issues of rail vehicle running smoothness are examined in the research, as one of the main subjects is rail vehicles' stability.

2. Methodology of research on running gear vibration

During the assessment of rail vehicle running gear vibration level and considering the passenger comfort, the SCI was used as an indicator of running smoothness [6, 29]. The value of SCI was calculated according to the formula:

$$W_Z = \left(\sum_{i=1}^{n_f} W_{Z_i}^{10} \right)^{\frac{1}{10}}, W_{Z_i} = \left[a_i^2 B(f_i)^2 \right]^{\frac{1}{6.67}}, \quad (1)$$

where: n_f - the number of frequencies considered, a - carbody acceleration, m/s^2 , f_i - vibration frequency, Hz, $B(f_i)$ - frequency and vibration direction coefficient influencing the passenger well-being:

$$B(f_i) = k \sqrt{\frac{1.911 f_i^2 + (0.25 f_i^2)^2}{(1 - 0.277 f_i^2)^2 + (1.563 f_i - 0.0368 f_i^3)^2}}, \quad (2)$$

where: $k = 0.737$, if oscillations are lateral, and $k = 0.588$ if oscillations are vertical.

The smooth-running indicators calculated based on Equations (1-2) are compared with the standard assessment scale. The quality of rail vehicle running gear behaviour is finally assessed according to comparative results.

At SCI values up to 1, the human senses do not feel the impact of vibrations; at SCI values from 1 to 3, vibrations are felt but do not cause any discomfort, and at SCI values from 3.0 to 3.5, the discomfort is felt. Exceeding the SCI value of more than 4, the vibrations are hazardous to human health. Therefore, the SCI limit for vehicles is taken up to 3.25.

Based on this methodology, examples of the values of the SCI under the specific operating conditions of a passenger car running gear are further analysed, for example, the case of a passenger car with independently rotating wheels, a small radius curve or when the wheel running surface has damage.

At first, the SCI was modelled for a passenger car with IRWs and with typical (unmodified) suspension, which parameters are presented in Table 1.

Table 1. Typical parameters for passenger car suspension

| Parameter | Value |
|---|----------------|
| Primary suspension stiffness coefficient in the lateral direction, N/m | $1 \cdot 10^6$ |
| Primary suspension stiffness coefficient in the vertical direction, N/m | $1 \cdot 10^6$ |
| Secondary suspension stiffness coefficient in the lateral direction, N/m | $2 \cdot 10^5$ |
| Secondary suspension stiffness coefficient in the vertical direction, N/m | $2 \cdot 10^5$ |
| Total damping factor of primary and secondary suspension, Ns/m | $1 \cdot 10^4$ |

3. Results of numerical modelling of a passenger car running

3.1. Sperling's comfort index values of typical suspension of passenger car

During the study, SCI in lateral and vertical directions at different running speeds were simulated by software package "Universal Mechanism" (UM) on different sections of the track. The obtained values of SCI are provided in Figure 1 and Figure 2.

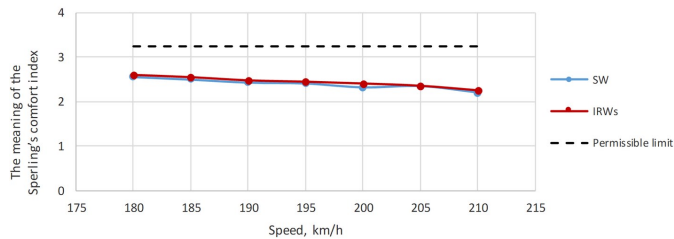


Fig. 1. Sperling's comfort index values in the vertical direction in the track tangent section

The diagram of Figure 1 shows that the SCI (in terms of vertical oscillations) decreases steadily with increasing speed from 180 km/h for both the one solid wheelset (SW) and the independently rotating wheels.

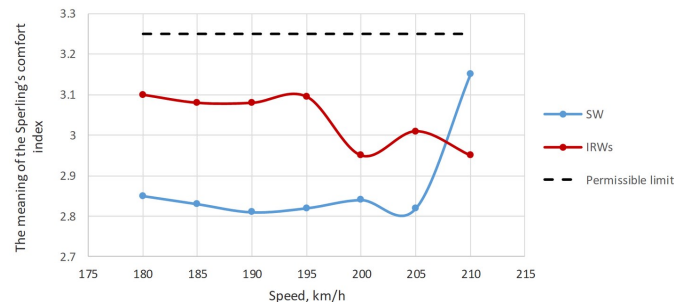


Fig. 2. Sperling's comfort index values in the lateral direction in the track tangent section

The curves of Figure 2 show that divergences occur in the variation of the SCI in terms of lateral vibrations at a speed of 200 km/h. Examining the change of the SCI according to the speed when the track section is tangent, different tendencies of the criterion change can be seen by analysing the oscillations in the vertical and horizontal planes. The values calculated from the vibration parameters of the vertical plane decrease gradually with increasing speed from 180 km/h to 210 km/h. Meanwhile, in the horizontal plane, at a speed of 200-210 km/h, divergences of value change are observed. The graphs of the variation of the values of the SCI in the 200 m radius curve according to the speed are shown in Figure 3 and Figure 4.

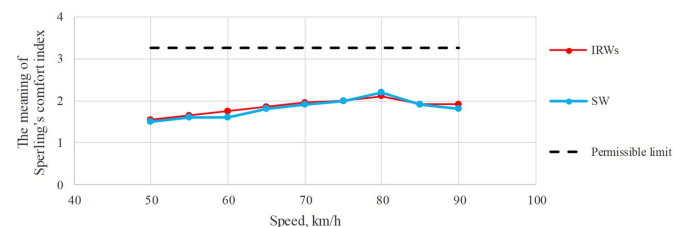


Fig. 3. Sperling's comfort index values in the vertical direction in 200 m radius curve

The diagram of Fig. 3 shows that the SCI changes consistently in terms of vertical oscillations, and in the 200 m radius curve, only the

lower speed range is considered in the curve; the SCI, in this case, increases consistently (almost consistently).

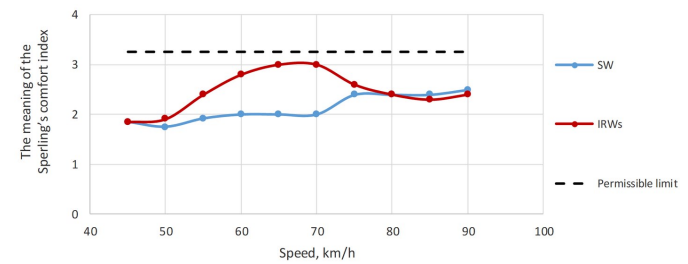


Fig. 4. Sperling's comfort index values in the lateral direction in 200 m radius curve

As in the tangent section of the track, divergences (in the speed range 60-70 km/h) are observed in the change of the SCI in terms of lateral oscillations with the 200 m radius curve (Fig. 4). Examining the change of the SCI in terms of the speed at the 200 m radius of the track curve, as in the case of a tangent track, different trends of the criterion change can be seen by analysing the oscillations in the vertical and lateral planes. The values calculated from the vertical plane oscillation parameters increase steadily as the speed increases from 50 km/h to 80 km/h (the trend changes slightly at 90 km/h). In the horizontal plane, at speeds of 60-70 km/h, the divergences of change of SCI values are observed. The Authors of the study pointed out that so far, only cases with standard passenger car suspension have been considered. By changing the stiffness of the rail vehicle suspension, the dynamic parameters of the passenger car running also change.

3.2. Sperling's comfort index values of adjusted suspension of passenger car

In order to improve the dynamic parameters of the passenger car with IRWs, the stiffness values of the primary and secondary suspension elements of their running gear were adjusted. Prior to adjusting the values, a study was performed to determine how the mean square of carbody accelerations depend on the stiffness of the respective suspension [28]. The dependences of the mean square of carbody accelerations on the stiffness of the primary suspension are presented in Figure 5 and Figure 6.

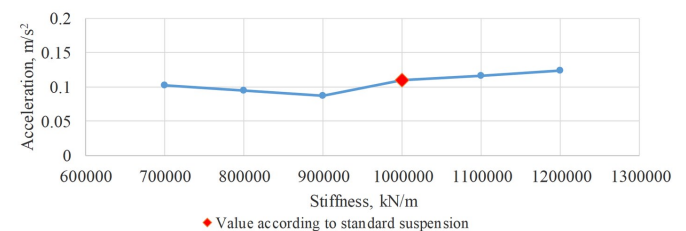


Fig. 5. Dependence of mean square of carbody accelerations on the vertical stiffness of the primary suspension

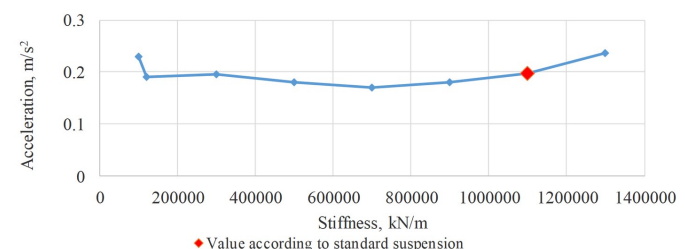


Fig. 6. Dependence of mean square of carbody accelerations on the lateral stiffness of the primary suspension

The dependences of the mean square of carbody accelerations on the vertical and the lateral stiffness of the primary suspension, respec-

tively, indicate that the stiffness of the standard suspension elements needs to be adjusted to improve the dynamic characteristics of the passenger car with IRWs.

The dependences of the mean square of carbody accelerations on the stiffness of the secondary suspension are presented in Figure 7 and Figure 8.

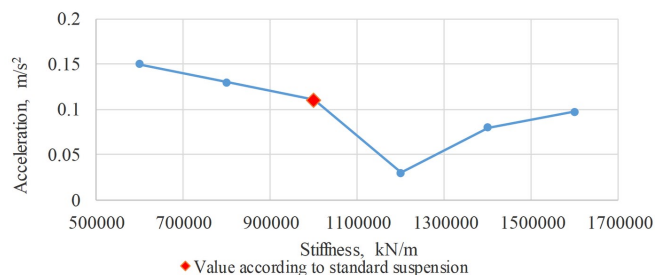


Fig. 7. Dependence of mean square of carbody accelerations on the vertical stiffness of the secondary suspension

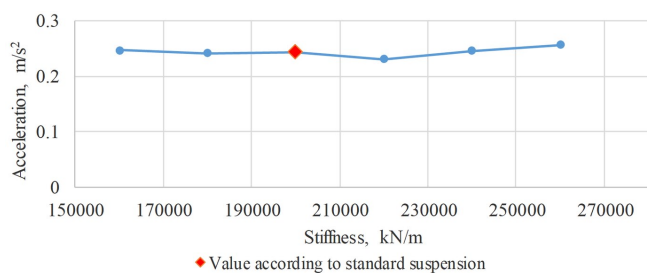


Fig. 8. Dependence of mean square of carbody accelerations on the lateral stiffness of the secondary suspension

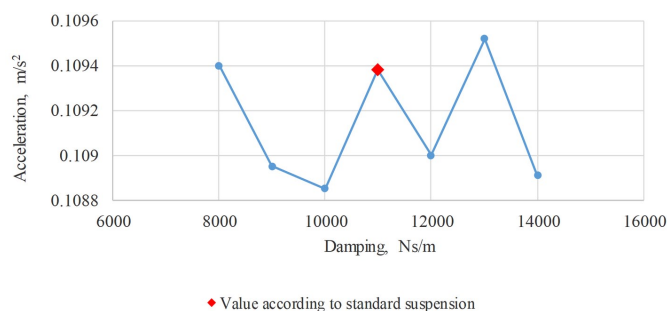


Fig. 9. Dependence of mean square body accelerations on secondary suspension damping parameters

The dependence of the mean square of carbody accelerations on the secondary suspension damping parameters is shown in Figure 9.

The dependence of the mean square of carbody accelerations of the secondary suspension for vertical and lateral stiffness, respectively, as well the dependence on the mean square carbody accelerations on the secondary suspension damping parameters indicate that the stiffness of the standard suspension elements also needs to be adjusted to improve the IRW dynamic performance.

Based on the research data, the stiffness values of the suspension elements were chosen. These data are submitted in Table 2.

By using the newly selected values of the stiffness of the passenger car suspension elements, the regularities of the change of SCI values were remodelled. SCI gained values are presented in Figures 10 and Figure 11, respectively.

The curves of Fig. 10 show that in the case of a standard suspension, the SCI on tangent track (in terms of vertical vibrations) decreases steadily with increasing the speed from 180 km/h for both the SW and the IRWs.

The curves of Figure 11 show that, as with the standard suspension, the SCI divergences occur on the tangent section in the variation of the SCI in terms of lateral vibrations. This is especially true in the case

Table 2. Values of stiffness coefficients of passenger car suspension elements after adjustment

| Parameter | Value |
|---|------------------|
| Primary suspension stiffness coefficient in the lateral direction, N/m | $9 \cdot 10^5$ |
| Primary suspension stiffness coefficient in the vertical direction, N/m | $9 \cdot 10^5$ |
| Secondary suspension stiffness coefficient in the lateral direction, N/m | $2.2 \cdot 10^5$ |
| Secondary suspension stiffness coefficient in the vertical direction, N/m | $1.2 \cdot 10^6$ |
| Total damping factor of primary and secondary suspension, Ns/m | $7 \cdot 10^4$ |

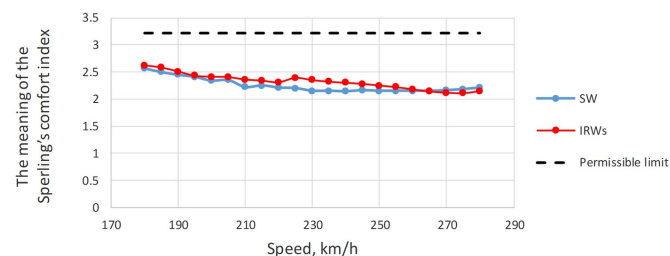


Fig. 10. The value of the Sperling's comfort index in the vertical direction in tangent track

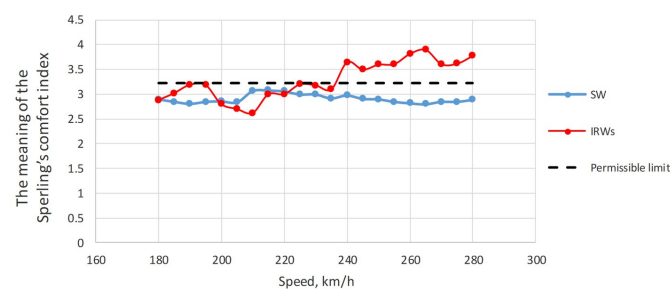


Fig. 11. The value of the Sperling's comfort index in the lateral direction in tangent track

of IRWs. Comparing the tendencies of the change of the value of the SCI on the track tangent section in the vertical and lateral directions was noticed that in the vertical direction, the consistent decrease is observed with increasing running speed. The divergences of change of SCI value in the lateral direction in the, are observed when the speed of a passenger car with IRWs reaches the values of (240-280) km/h.

The variation of SCI values on the 200 m radius curve is shown in Figure 12 and Figure 13.

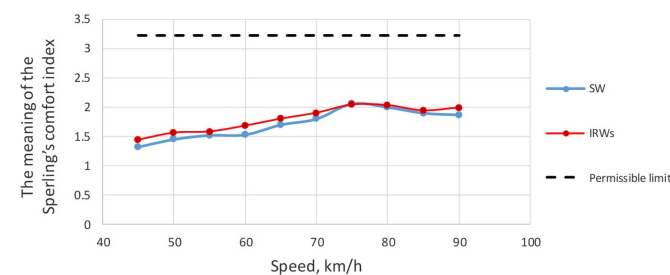


Fig. 12. The value of the Sperling's comfort index in the vertical direction in a 200 m radius curve

SCI values in the vertical direction on a track curve with a radius of 200 m, as in the case of a standard suspension, changes consistently, i.e. without observable divergences.

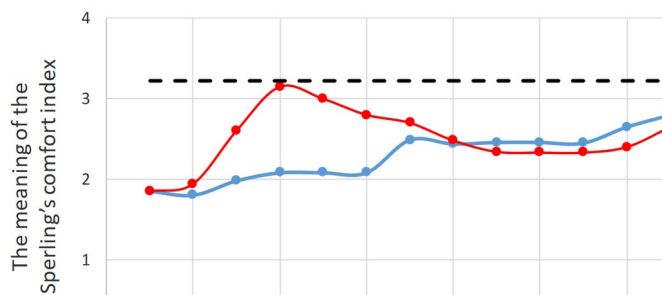


Fig. 13. The value of the Sperling's comfort index in the lateral direction on a curve with a radius of 200 m

The divergences of the variation of SCI values in the lateral direction on the 200 m radius curve of the track, as in the case of standard suspension, were observed. The change of SCI values according to the speed in the 200 m radius curve shows the same tendencies as previously analysed: in the lateral direction – consistent change, in the lateral direction – divergences appear.

The modelling results show that the divergences of the change of the SCI occur precisely due to the oscillations in the lateral direction. This fact raises the question to the study Authors as to whether this is not a systematic error in the modelling (e.g., the assumption made in the programmed conditions). For searching for an answer to this question, the Authors of this study conducted further study, which included not only the theoretical calculation of the SCI but also its determination based on the vibration parameters measured in field tests. A specific case of observation was a passenger car with a damaged wheel.

3.3. Sperling's comfort index values in case of a damaged wheel

As in the other cases examined, the SCI in the presence of a damaged wheel was primarily modelled by numerical simulation. These wheel damage parameters for the modelling were selected: flat depth $h = 0.001$ m and length $L = 20$ mm.

After having processed the data obtained during the simulation by means of the Fourier transform method, the dependence of the vehicle body acceleration amplitude repetitions on the running time was obtained. These data make it possible to assess the comfort of the passengers through the SCI. The obtained SCI values are shown in Figure 14.

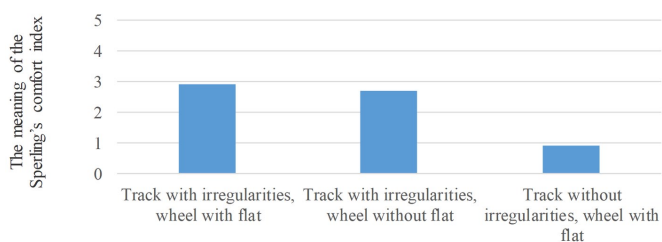


Fig. 14. Values of Sperling's comfort index

The diagram of Figure 14 shows that the SCI values of running smoothness are acceptable (see Table 2) when the passenger cars operate with the damaged wheel with a flat depth of 1 mm depth and a length of 20 mm on the track tangent section. However, after introducing track roughness, the SCI values increased about 3 times and approached the limit values for passenger cars.

To test these data and the previously hypothesised that the divergences of the change of the SCI values occur when examining the oscillations in the lateral direction of the passenger car, a field test (experiment) was performed. During it, the oscillations of the passen-

ger carbody were measured in practice, and the trends of SCI value change were determined based on the results.

4. Identification of passenger car running smoothness parameters by field testing

The main parameters of the measurement equipment used for the experiment are presented in Table 3. The mounting of the sensors in the passenger car is shown in Figure 15. The recorded data of the experiment were estimated by the SCI for the assessment of smoothness of passenger car rides, and gained results are presented in Figure 16.

As seen from Figure 16, the SCI values comply with the requirements for ride stability, sufficient for passenger cars, whereas good re-

Table 3. Basic parameters of the equipment used for the experiment

| Equipment | Measurement limits, g | Measurement frequency, Hz | Sensor mass, kg | Accuracy, % |
|------------------------|-----------------------|---------------------------|-----------------|-------------|
| Corrsys-Datron HF-500C | ± 3 | 10 | 0.230 | ± 0.2 |
| Kistler Type 8395A | ± 3 | 1000 | 0.155 | ± 0.2 |



Fig. 15. Mounting of the sensors in the passenger car

sults have not been reached in the lateral direction. The highest value of the index in lateral direction has been reached at the running speed of 40 km/h, while with speed increasing, the index values went on decreasing. The SCI values received in the vertical direction fluctuates in the zone of "sufficient for passenger cars". With the car running speed augmenting, they are evenly increasing.

The insight of the experiment is that SCI values in terms of vertical oscillations change consistently with varying speeds. In contrast, the regularity of the change of these values in the lateral direction deviates from the consistent change. Therefore, it is expedient to analyse the regularity of the change of the values of the SCI according to the vibrations in the lateral direction.

The experimentally determined regularity of the change of the SCI (according to the oscillations in the lateral direction) can be described by the equation of the 2nd, the 3rd or the 4th degree, respectively, the following expressions are possible:

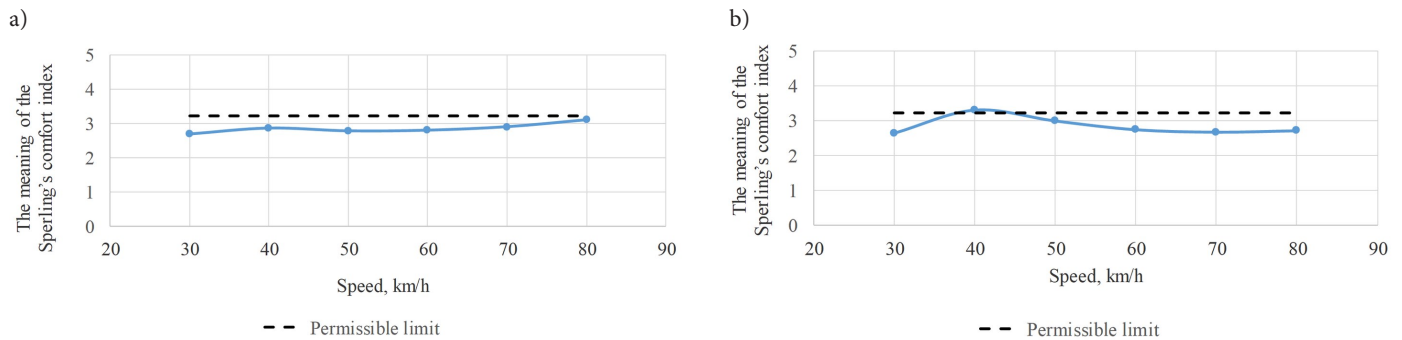


Fig. 16. Experimental Sperling's comfort index values at different running speeds: a) in vertical direction; b) in lateral direction

$$\begin{aligned} f(x) &= p_1x^2 + p_2x + p_3; \\ f(x) &= p_1x^3 + p_2x^2 + p_3x + p_4; \\ f(x) &= p_1x^4 + p_2x^3 + p_3x^2 + p_4x + p_5. \end{aligned} \quad (3)$$

The values of the coefficients p_i of the 2nd, the 3rd and the 4th-degree function (coefficients 3, 4 and 5, respectively) and the coefficients of determination R^2 , respectively, are given in Table 4.

Table 4. Coefficients of polynomial function according to experimental data

| Coefficients | Values | | |
|--------------|-------------------------------------|-------------------------------------|-------------------------------------|
| | The 2 nd degree function | The 3 rd degree function | The 4 th degree function |
| p_1 | - 0.0004 | 0.00005 | - 0.000002 |
| p_2 | 0.0369 | - 0.0092 | 0.0006 |
| p_3 | 2.0884 | 0.494 | - 0.0492 |
| p_4 | - | - 5.304 | 1.848 |
| p_5 | - | - | - 21.69 |
| R^2 | 0.304 | 0.8628 | 0.9938 |

According to the last row of Table 4, the coefficient of determination of the 2nd-degree mathematical correlation is too small: $R^2 = 0.304$, the coefficient of determination of the 3rd-degree function $R^2 = 0.8628$, and the coefficient of determination of the 4th-degree function $R^2 = 0.9938$ – very strong mathematical correlation. It can be concluded that it is recommended to describe the regularity of the variation of SCI according to the rail vehicle speed with the equation (polynomial) of the 3rd or the 4th degree. The accuracy of SCI description ensures the possibility to maintain an acceptable level of a passenger car running smoothness during exploitation.

5. Conclusions

The investigation of the tendency of Sperling's comfort index variation is an appropriate way to define the smoothness of passenger cars under various exploitation conditions. By monitoring the variation of the Sperling's comfort index according to the running speed of the vehicle, ride smoothness level can be assessed. In cases such as when

vehicle running gear is with independently rotating wheels, when a car is curving the small radius curves of a track or when the wheel surfaces are damaged.

To verify the suitability of Sperling's comfort index for assessing the smoothness of a passenger car ride, the Authors performed a numerical simulation of the vehicle with independently rotating wheels. These simulations are running in a track tangent section and 200 m radius track curve and adjusting the passenger car suspension parameters.

The Authors also performed a numerical simulation and experimental research of a passenger car with a damaged wheel running surface and compared the obtained results. During examining the variation of the Sperling's comfort index according to the running speed of the vehicle, it was observed that to operate the vehicles safely with independently rotating wheels, and it is necessary to adjust the suspension parameters of the passenger car. It is necessary to change the stiffness and damping of the primary and secondary suspension parameters. Otherwise, Sperling's comfort index values, especially in the lateral direction, change chaotically and indicate the inadmissible quality of ride smoothness control.

After summarising the results of this study, the Authors recommend describing the tendency of the variation of the Sperling's comfort index (considering the running gear oscillations in the lateral direction) according to the rail vehicle speed, with the equations of the 3rd or the 4th-degree. The coefficient of determination was defined by describing the dependence of the 4th-degree function ($R^2=0.9938$ – a very strong mathematical correlation).

The investigation of vibration parameters of passenger cars shows that the divergences of the change of the Sperling's comfort index occur especially due to the oscillations in the lateral direction – this is confirmed both by theoretical calculations and by vibration parameters measured in practice. Therefore, the study of the regularities of lateral oscillations of a passenger car is a reasonable direction for further research.

Acknowledgment

The research is funded by the EU Shift2Rail project GEARBODIES (Grand number: 10101396) under Horizon 2020/ Shift2Rail Framework Programme.

References

- Brommundt E. A simple mechanism for the polygonalization of railway wheels by wear. *Mechanics Research Communications and Applied* 1997; 24(2): 435-442, [https://doi.org/10.1016/S0093-6413\(97\)00047-5](https://doi.org/10.1016/S0093-6413(97)00047-5).
- Bureika G, Levinzon M, Dailydka S, Steišūnas S, Žygienė R. Evaluation criteria of wheel/rail interaction measurement results by trackside control equipment. *International Journal of Heavy Vehicle Systems* 2019; 26(6): 747-764, <https://doi.org/10.1504/IJHVS.2019.102682>.
- Dukalski P, Będkowski B, Parczewski K, Wnęk H, Urbaś A, Augustynek K. Dynamics of the vehicle rear suspension system with electric motors mounted in wheels. *Eksplotacja i Niezawodność - Maintenance and Reliability* 2019; 21(1): 125-136, <https://doi.org/10.17531/ein.2019.1.14>.
- Ekberg A, Kabo E. Fatigue of railway wheels and rails under rolling contact and thermal loading-an overview. *Wear* 2005; 258(8): 1288-1300, <https://doi.org/10.1016/j.wear.2004.03.039>.

5. Favorskaya A, Khokhlov N. Modelling the impact of wheelsets with flat spots on a railway track. *Procedia Computer Science* 2018; 126: 1100-1109, <https://doi.org/10.1016/j.procs.2018.08.047>.
6. Frischmuth K, Langemann D. Numerical Analysis of long-term wear models. *Machine Dynamics Problems* 1998; 20: 113-122.
7. Grassie S, Gregory R, Harrison D, Johnson K. The dynamic response of railway track to high frequency vertical, lateral, longitudinal excitation. *Journal of Mechanical Engineering Science* 1982; 24(2): 77-102, https://doi.org/10.1243/JMES_JOUR_1982_024_016_02.
8. Gorbunov M, Kravchenko K, Bureika G, Gerlici J, Nozhenko O, Vaičiūnas G, Bučinskas V, Steišūnas S. Estimation of sand electrification influence on locomotive wheel/ rail adhesion processes. *Eksplotacija i Niezawodnosc - Maintenance and Reliability* 2019; 21(3): 460-467, <https://doi.org/10.17531/ein.2019.3.12>.
9. Hayes W, Tucker H. Wheelset-track resonance as a possible source of corrugation wear. *Wear* 1991; 144(1-2): 211-226, [https://doi.org/10.1016/0043-1648\(91\)90016-N](https://doi.org/10.1016/0043-1648(91)90016-N).
10. Jiang Y, Bernard K, Chen B. K, Thompson C. A comparison study of ride comfort indices between Sperling's method and EN12299. *International Journal of Rail Transportation* 2019; 7(4): 1-18, <https://doi.org/10.1080/23248378.2019.1616329>.
11. Jin X, Xiao X, Wen Z, Guo J, Zhu M. An investigation into the effect of train curving on wear and contact stresses of wheel and rail. *Tribology international* 2009; 42(3): 475-490, <https://doi.org/10.1016/j.triboint.2008.08.004>.
12. Kim Y, Kwon H, Kim S. Correlation of ride comfort evaluation methods for railway vehicles. *International Journal of Rail Transportation* 2003; 17(217): 73-88, <https://doi.org/10.1243/095440903765762823>.
13. Kowalski S. The influence of selected PVd coatings on fretting wear in a clamped joint based on the example of a rail vehicle wheel set. *Eksplotacija i Niezawodnosc - Maintenance and Reliability* 2018; 20(1): 1-8, <https://doi.org/10.17531/ein.2018.1.1>.
14. Kufver B, Persson R, Wingren J. Certain aspects of the CEN standard for the evaluation of ride comfort for rail passengers. *WIT Transactions on The Built Environment* 2010; 29: 605-614, <https://doi.org/10.2495/CR100561>.
15. Kusel M, Brommundt E. The evolution of noncircularities at braked or driven railway wheels. *Machine Dynamics Problems* 1998; 20: 313-324.
16. Ma Ch, Gao L, Xin T, Cai X, Nadakatti M, Wang P. The dynamic resonance under multiple flexible wheelset-rail interactions and its influence on rail corrugation for high-speed railway. *Journal of Sound and Vibration* 2021; 498, <https://doi.org/10.1016/j.jsv.2021.115968>.
17. Meywerk M. Polygonalization of railway wheels. *Archive of Applied Mechanics* 1999; 69(2): 105-120, <https://doi.org/10.1007/s004190050208>.
18. Michitsuji Y, Suda Y. Running performance of power-steering railway bogie with independently rotating wheels. *Vehicle System Dynamics* 2006; 4(1): 71-82, <https://doi.org/10.1080/00423110600867416>.
19. Munawir T, Samah A, Rosle M. A comparison study on the assessment of ride comfort for LRT passengers. *International Research and Innovation Summit (IRIS2017)*, 2017, <https://doi.org/10.1088/1757-899X/226/1/012039>.
20. Peng B, Iwnicki S, Shackleton Ph, Crosbee D, Zhao Y. The influence of wheelset flexibility on polygonal wear of locomotive wheels. *Wear* 2019; 432-433, <https://doi.org/10.1016/j.wear.2019.05.032>.
21. Perez J, Mauer M, Busturia J. M. Design of Active Steering Systems for Bogie-Based Railway Vehicles with Independently Rotating Wheels. *Vehicle System Dynamics* 2002; 37(1): 209-220, <https://doi.org/10.1080/00423114.2002.11666233>.
22. Piotrowski J. A substitute model of two-dimensional dry friction exposed to dither generated by rolling contact of wheel and rail. *Vehicle System Dynamics* 201; 50(10): 1495-1514, <https://doi.org/10.1080/00423114.2012.676653>.
23. Polach O. Wheel profile design for the target conicity and wide tread wear spreading. *Wear* 2011; 271(1-2): 195-202, <https://doi.org/10.1016/j.wear.2010.10.055>.
24. Popp K, Kruse H, Kaiser I. Vehicle-track dynamics in the mid-frequency range. *Vehicle system dynamics C. International Journal of Vehicle Mechanics and Mobility* 1999; 31(5-6): 423-464, <https://doi.org/10.1076/vesd.31.5.423.8363>.
25. Pradhan S, Samantaray A, Bhattacharyya R. Evaluation of ride comfort in a railway passenger vehicle with integrated vehicle and human body bond graph model. *International Mechanical Engineering Congress and Exposition*, 2017, <https://doi.org/10.1115/IMECE2017-71288>.
26. Sladkowski A, Pogorelov D. Investigation of the dynamic interaction in the wheel-rail contact in the presence of flat spots on the wheelset. *Bulletin of the East Ukrainian National University* 2008; 5(123): 88-95.
27. Suda Y, Wang W, Nishina M, Lin S, Michitsuji Y. Self-steering ability of the proposed new concept of independently rotating wheels using inverse tread conicity. *Vehicle System Dynamics* 2012; 50(1): 291-302, <https://doi.org/10.1080/00423114.2012.672749>.
28. Taletavičius, R. Rolling stock driving stability study. Master thesis, VGTU, 2019: 73.
29. Vaičiūnas G, Steišūnas S. Sperling's comfort index study in a passenger car with independently rotating wheels. *Transport Problems* 2021, 16(2): 121-130.
30. Vaičiūnas G, Bureika G, Steišūnas S. Research on metal fatigue of rail vehicle wheel considering the wear intensity of rolling surface. *Eksplotacija i Niezawodnosc - Maintenance and Reliability* 2018, 20(1): 24-29, <https://doi.org/10.17531/ein.2018.1.4>.
31. Wallentin M, Bjarnhed H, Lundén R. Cracks around railway wheel flats exposed to rolling contact loads and residual stresses. *Wear* 2005; 258(7-8): 1319-1329, <https://doi.org/10.1016/j.wear.2004.03.041>.
32. Ye Y, Sun Y, Shi D, Peng B, Hecht M. A wheel wear prediction model of non-Hertzian wheel-rail contact considering wheelset yaw: Comparison between simulated and field test results. *Wear* 2021; 474-475, <https://doi.org/10.1016/j.wear.2021.203715>.

The post-warranty random maintenance policies for the product with random working cycles

Lijun Shang^a, Haibin Wang^{b,c}, Cang Wu^c, Zhiqiang Cai^{b,*}

^aFoshan University, School of Quality Management and Standardization, Foshan 528225, China

^bNorthwestern Polytechnical University, School of Mechanical Engineering, Xi'an 710072, China

^cChina United Northwest Institute for Engineering Design & Research Co., Ltd., Xi'an 710077, China

^dLanzhou University of Technology, School of Mechanical and Electrical Engineering, Lanzhou 730050, China

Indexed by:



Highlights


- A novel warranty of the product with random working cycles is proposed.
- Random maintenance policies sustaining the post-warranty reliability are investigated.
- The performance of replacement last (first) with PM is more excellent.

Abstract

Advanced sensors and measuring technologies make it possible to monitor the product working cycle. This means the manufacturer's warranty to ensure reliability performance can be designed by monitoring the product working cycle and the consumer's post-warranty maintenance to sustain the post-warranty reliability can be modeled by tracking the product working cycle. However, the related works appear seldom in existing literature. In this article, we incorporate random working cycle into warranty and propose a novel warranty ensuring reliability performance of the product with random working cycles. By extending the proposed warranty to the post-warranty maintenance, besides we investigate the post-warranty random maintenance policies sustaining the post-warranty reliability, i.e., replacement last (first) with preventive maintenance (PM). The cost rate is constructed for each post-warranty random maintenance policy. Finally, sensitivity of proposed warranty and investigated policies is analyzed. We discover that replacement last (first) with PM is superior to replacement last (first).

Keywords

random working cycle, warranty, post-warranty reliability, replacement last, replacement first.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

1. Introduction

In a broader sense, offering products to warranty can benefit simultaneously manufacturers and consumers. Due to this, the warranty policies (or models) have been recently researched widely from the manufacturer's perspective. The research stream on warranty policies is dependent on reliability modeling (or evaluation) technology. It is less difficult to discover that there are two types of reliability modeling technology, which are being used frequently in academia and industry. One type is that the product lifetime is modeled as a distribution function with self-announcing failure. For example, Li et al. [11] studied the reliability evaluation using limited and censored time-to-failure data, by means of the uncertainty theory; Zhang and Zhang [29] first proposed a reliability model of aviation cables by using nonlinear mixed model and Bayesian estimation, and then analyzed accuracy of reliability model by the failure time of cable. Other type is that the product failure is modeled as a type of degradation failure (which is referred to as not self-announcing failure [18]). For example, Gao et al. [5] developed methods to analyze the system reliability

of two-phase degradation model with a random change point; Huang et al. [6] proposed a degradation model for soft failure by considering continuous degradation processes with recoverable shock damages for reliability assessment and lifetime prediction of products.

Along with the above frame on reliability modeling technology, similarly, the research stream on warranty policies can be distinctly divided into two research streams. The first stream concentrates on the design of the distribution-based warranty policies, namely design warranty policies by modeling the product lifetime as a distribution function with self-announcing failure. For example, Hooti et al. [7] proposed an extended two-dimensional warranty plan which includes limitation on time and the number of repairs, under the assumption that the lifetime of the system follows distribution function; Huang et al. [8] developed a model to determine the optimal sale price, warranty period and product reliability to maximize the discounted profit for a repairable product sold with a free replace-repair warranty policy, by assuming that the product failure time follows distribution function; He et al. [9] established the decision model of extended warranty price from the perspective of win-win by assuming that the product

(*) Corresponding author.

E-mail addresses: L. Shang - ljshang2020@126.com, H. Wang - 36185397@qq.com, C. Wu - 464968118@qq.com, Z. Cai - caizhiqiang@nwpu.edu.cn

failure time follows distribution function; Knopik and Migawa [10] investigated the effects of introducing preventive replacement to maintenance system implemented by age-replacement of technical objects with valid manufacturer's warranty and non-repairable, by means of Weibull distribution. Wang et al. [24] studied an optimal extended warranty policy after the expiration of base two-dimensional warranty with repair time threshold by assuming that the failure time of the equipment follows distribution function. The second stream aims to design the degradation-based warranty policies, namely design warranty policies by modeling the product failure as a type of degradation failure. For example, Cha et al. [2] and Zhang et al. [30] studied warranty policy of the product by modeling the product failure as degradation failure.

Usually, the manufacturer adopts some methods (maintenance or replacement, and so on) to ensure the product reliability performance during the warranty period. However, consumers (or users) tend to be concerned about how to sustain reliability during the post-warranty period (i.e., the post-warranty reliability). Due to increased maintenance costs, how to model a post-warranty maintenance to sustain the post-warranty reliability has recently received considerable attention. This type of problem has been also investigated extensively along with the above reliability modeling technology. For example, Liu et al. [14] investigated the optimal replacement problem for a warranty product subject to $(M + 1)$ types of mutually exclusive failure modes, including M repairable failure modes and a catastrophic failure mode, by supposing that the warranty product's lifetime follows Weibull distribution; Park et al. [17] developed mathematical formulas to evaluate the long-run expected cost rates during the life cycle of the product, by considering the failure time of the product and a Weibull distribution; and Shang et al. [19] investigated an optimal maintenance-replacement policy after the warranty expiry by assuming that the product lifetime follows distribution function; Shang et al. [20] investigated the post-warranty maintenance by modeling the product failure as a degradation failure.

By modeling the product failure as a type of degradation failure, in essence, designing warranty policies and modeling post-warranty maintenance are undoubtedly driven and powered by in-situ sensor and measuring technologies, which can accurately or approximately measure/inspect health condition of the product. In addition to measuring product health condition, in-situ sensor and measuring technologies can monitor working cycle of the product effectively that performs successively projects or missions at random working cycle. In real world, lots of products work at random working cycle. For example, the intelligent air pump inflates the tire at random working cycle; and the intelligent cutter cuts the material at random working cycle, and so on. For the product with random working cycles (i.e., the product which works at random working cycle), from reliability theory, it deteriorates with respect to its working time. Considering this reality, Chang [3] and Sheu et al. [21, 22] researched preventive maintenance policies to ensure or enhance the product reliability performance of the product with random working cycles by modeling the product working cycles as an independent identically distributed random variable sequence; Nakagawa [16] and Zhao et al. [31] researched various maintenance policies of the product with random working cycles by assuming working cycle as an independent identically distributed random variable.

If the product working cycle is integrated into the warranty period and the post-warranty period, the following advantages can be brought: ① the manufacturer or the consumer can calculate the product reliability by making use of the monitored total working time in real time; ② The manufacturer can more precisely evaluate the warranty budget and more efficiently control warranty cost; ③ the post-warranty maintenance planning techniques (such as repair, replacement, imperfect preventive maintenance) of the consumer can be programmed and scheduled more reasonably, and the related maintenance cost can be reduced appropriately, and so on. However, few warranty policies and the post-warranty maintenance policies to sus-

tain the post-warranty reliability have been developed by integrating the product working cycle.

In this article, we introduce the limited number of random working cycle to the warranty period and proposes a novel warranty from the manufacturer's perspective. The proposed warranty requires that if the failure doesn't occur until the warranty period before the completion of the limited number of random working cycle, then the proposed warranty expires at the warranty period; and if the failure doesn't occur until the completion of the limited number of random working cycle before the warranty period, then free repair (minimal repair) warranty [4, 15] will be triggered to warrant the product from the completion of the limited number of random working cycle to the warranty period. Defining that the proposed warranty is extended to the consumer's post-warranty maintenance model, we investigate two kinds of the post-warranty random maintenance policy to sustain the post-warranty reliability. The first type is replacement last with preventive maintenance (PM), where PM at the warranty period is integrated into random periodic replacement last [16]. The second type is replacement first with PM where PM at the warranty period is integrated into random periodic replacement first [16]. For each post-warranty random maintenance policy, we construct the related cost rate model by integrating the product's depreciation expense depending on the total working time. By means of the numerical experiments, we compare the performance of the post-warranty random maintenance policies.

The contribution of this article can be highlighted in three key aspects: (1) we propose a novel warranty to ensure reliability performance of the product with random working cycles; (2) we investigate two types of random maintenance policy to sustain the post-warranty reliability of the product, which seldom exists in literature; (3) the performance of replacement last (first) with PM is more excellent.

The structure of this article is organized as follows. Section 2 proposes the manufacturer's warranty, derives the related warranty cost. In Section 3, replacement last with PM and replacement first with PM are defined, and the related cost rate models are derived. Section 4 presents a comparing approach, which can help manufacturer to make decision on the post-warranty random maintenance policies. In Section 5, numerical experiments are used to illustrate the proposed approach and sensitivity analysis on some key parameters is performed. Finally, conclusions are drawn in Section 6.

2. Warranty model

It is assumed that the product does projects or missions successively, and random working cycle Y_j of the j^{th} ($j = 1, 2, \dots$) project is independent identically distributed to the distribution function $G(y) = \Pr\{Y_j < y\}$ with the lack-of-memory property. The product deteriorates with respect to its working time and the time-to-first-failure X of the product is subject to a general distribution function $F(x) = \Pr\{X < x\}$ with a failure rate function $r(u)$ where $u > 0$. Besides, it is assumed that the downtime resulted from each replacement or each minimal repair is completely negligible in this article.

2.1. Warranty assumptions

The particular attractiveness of renewing (or renewable) free replacement warranty [13, 19] (RFRW) is that it makes possibly consumers to obtain freely a new identical product with the same warranty. Due to this particular characteristic, RFRW is an attractive warranty which can be used as a significant advertising tool from the manufacturer's perspective. Basing on the product working cycle monitored by using in-situ sensor and measuring technologies, besides the manufacturer can design warranty policies to ensure the product reliability performance. However, the existing RFRW model neglects universally to make proper use of the product working cycle.

In view of this, we consider the particular attractiveness of RFRW and study a novel warranty of the product which performs successively projects or missions at random working cycle, as below.

Given the number m (a limited value) of random working cycle and the warranty period w , the warranty proposed in this article is described as follows:

- (1) The product will be replaced by a new identical one with the warranty proposed in this section (i.e., failure replacement) if the failure occurs before the completion of the m^{th} random working cycle or before the warranty period w , whichever occurs first.
- (2) The manufacturer shoulders whole failure replacement cost (including labor cost, transport cost and so on) resulted from unit failure replacement.
- (3) If the failure doesn't occur until the completion of the m^{th} random working cycle before the warranty period w , then free repair warranty [24, 28] (FRW) with a time span $w - S_m$ will be triggered to warrant the product, where S_m is the total working time of the product when the m^{th} random working cycle is completed and satisfies $S_m = \sum_{j=1}^m Y_j$; if the failure doesn't occur until the warranty period w before the completion of the m^{th} random working cycle, then the warranty expires.
- (4) FRW requires that any failure in the interval $(S_m, w]$ is removed by minimal repair and the related minimal repair cost is also shouldered by the manufacturer.

Note that ① in this warranty, m and w are obviously two types of failure replacement limit, and so the warranty region related to failure replacement can be represented as $(0, m] \times (0, w]$; ② the product goes through warranty (i.e., the proposed warranty, hereinafter similarly) at the completion of the m^{th} random working cycle before the warranty period w or at the warranty period w before the completion of the m^{th} random working cycle, whichever occurs first; ③ for some consumers with a higher product working frequency, their warranty expires very easily at the completion of the m^{th} random working cycle before the warranty period w , therefore the manufacturer offers them a FRW so that they are treated as equal as other consumers whose warranty expires at the warranty period w before the completion of the m^{th} random working cycle.

2.2. Warranty cost modeling

Let $\bar{F}(x)$ be survival function of the time-to-first-failure X of the product, where $\bar{F}(x) = 1 - F(x)$. And let the Stieltjes convolution $G^{(m)}(s)$ ($G^{(m)}(s) = \int_0^s G^{(m-1)}(s-u)dG(u)$) and the Stieltjes convolution $\bar{G}^{(m)}(s)$ ($\bar{G}^{(m)}(s) = 1 - G^{(m)}(s)$) be respectively distribution function and survival function corresponding to the total working time S_m . According to the warranty proposed in Subsection 2.1, the case that the product goes through warranty can be divided into two types of case. The first case is that the product goes through warranty at the completion of the m^{th} random working cycle before the warranty period w ; and the second case is that the product goes through warranty at the warranty period w before the completion of the m^{th} random working cycle. By summing the probability of two types of case, the probability q that the product goes through warranty can be computed as:

$$q = \Pr\{S_m < w, S_m < X\} + \Pr\{w < S_m, w < X\} = \int_0^w \bar{F}(u)dG^{(m)}(u) + \bar{G}^{(m)}(w)\bar{F}(w) = 1 - \int_0^w \bar{G}^{(m)}(u)dF(u) \quad (1)$$

Since the event that the product doesn't go through warranty and the event that the product goes through warranty form jointly a complete event group, the probability p that the product doesn't go through warranty is expressed as:

$$p = 1 - q = \int_0^w \bar{G}^{(m)}(u)dF(u) \quad (2)$$

It is less difficult to conclude that the probability that until the i^{th} ($i=1, 2, \dots$) product goes through the m^{th} random working cycle or the warranty period w is a geometric distribution $p^{i-1}q$, and the number of failure replacement is precisely $i-1$. Further, the expected number of all failure replacements produced by the proposed warranty can be modeled as:

$$E[\kappa] = \sum_{i=1}^{\infty} p^{i-1}q(i-1) = \frac{p}{q} = \frac{\int_0^w \bar{G}^{(m)}(u)dF(u)}{1 - \int_0^w \bar{G}^{(m)}(u)dF(u)} \quad (3)$$

Let X_k ($k=1, 2, \dots$) be lifetime of the k^{th} product failed during the warranty region $(0, m] \times (0, w]$. According to probability theory, then every element of the sequence $\{X_k\}$ is independent identically distributed to the distribution function $H(x)$ with the below expression:

$$H(x) = \Pr\{X_k < x | X_k < S_m, X_k < w\} = \frac{\int_0^x \bar{G}^{(m)}(u)dF(u)}{\int_0^w \bar{G}^{(m)}(u)dF(u)} \quad (4)$$

where $0 < x < w$.

Suppose that the depreciation expense of the product is only affected by its working time t and is increasing with respect to t , then we model the depreciation expense $D(t)$ at t as:

$$D(t) = \alpha_1 t^{\beta_1} \quad (5)$$

where α_1 ($\alpha_1 > 0$) is depreciation rate; $0 < \beta_1 \leq \log_w(c_R / \alpha_1)$ where c_R is unit failure replacement cost suffered for the manufacturer.

For the k^{th} product failed during the warranty region $(0, m] \times (0, w]$, its working time is its lifetime X_k . So, the related depreciation expense is $D(X_k)$ and the k^{th} product failed prompts the manufacturer to suffer a cost $c_R - D(X_k)$. Until the $(i-1)^{\text{th}}$ failure replacement is completed, the replacement cost WC_{i-1} of the manufacturer can be obtained as:

$$WC_{i-1} = \sum_{k=0}^{i-1} (c_R - D(X_k)) \quad (6)$$

where $X_0 = 0$.

Since the random variable X_k is an independent and identically distributed to $H(x)$ in (4) and the number $i-1$ of failure replacement satisfies the geometric distribution $p^{i-1}q$, the expected value $E[WC_R]$ of the replacement cost WC_{i-1} can be obtained as:

$$E[WC_R] = E\left[\sum_{i=1}^{\infty} p^{i-1}q \cdot WC_{i-1}\right] = E\left[\sum_{i=1}^{\infty} p^{i-1}q \left(\sum_{k=0}^{i-1} (c_R - D(X_k))\right)\right] = \frac{p}{q} \cdot E[c_R - D(X_k)] = \frac{\int_0^w (c_R - D(x))\bar{G}^{(m)}(x)dF(x)}{1 - \int_0^w \bar{G}^{(m)}(u)dF(u)} \quad (7)$$

When the product goes through warranty at the completion of the m^{th} random working cycle, the total working time of the product is S_m . In this case, the distribution function $H_{S_m}(s)$ of the total working time S_m can be derived as:

$$H_{S_m}(s) = \Pr\{S_m < s | S_m < w, S_m < X\} = \frac{\int_0^s \bar{F}(u)dG^{(m)}(u)}{\int_0^w \bar{F}(u)dG^{(m)}(u)} \quad (8)$$

where $0 < s < w$.

By the third term [i.e., (3)] of the proposed warranty, when the product goes through warranty at the m^{th} random working cycle before the warranty period w , its past age is equal to its total working time S_m and it is warranted by FRW with a time span $w - S_m$. Let c_m be unit minimal repair cost, then the minimal repair cost WC_m produced by FRW can be estimated as:

$$WC_m = c_m \int_0^w r(S_m + u) du \quad (9)$$

Since the past age (i.e., the total working time) S_m is subject to the distribution function $H_{S_m}(s)$ in (8), the expected value $E[WC_m]$ of the minimal repair cost WC_m can be computed as:

$$E[WC_m] = E\left[c_m \int_0^w r(S_m + u) du\right] = c_m \int_0^w \left(\int_0^w r(s + u) du\right) dH_{S_m}(s) = \frac{c_m \int_0^w \left(\int_0^w r(s + u) du\right) \bar{F}(s) dG^{(m)}(s)}{\int_0^w \bar{F}(u) dG^{(m)}(u)} \quad (10)$$

It is well known that the probability that until the first product goes through the m^{th} random working cycle is $\sum_{i=1}^{\infty} p^{i-1} q_1 = q_1 / q$, where q_1 is the probability that the product goes through the m^{th} random working cycle and satisfies $q_1 = \Pr\{S_m < w, S_m < X\} = \int_0^w \bar{F}(u) dG^{(m)}(u)$; q has been offered in (1). By summing, the warranty cost $E[WC]$ produced by the proposed warranty can be derived as:

$$E[WC] = E[WC_R] + \frac{q_1}{q} \cdot E[WC_m] = \frac{\int_0^w (c_R - D(x)) \bar{G}^{(m)}(x) dF(x)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + \frac{\int_0^w \bar{F}(s) dG^{(m)}(s)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} \cdot \frac{c_m \int_0^w \left(\int_0^w r(s + u) du\right) \bar{F}(s) dG^{(m)}(s)}{\int_0^w \bar{F}(u) dG^{(m)}(u)} \\ = \frac{\int_0^w (c_R - D(x)) \bar{G}^{(m)}(x) dF(x) + c_m \int_0^w \left(\int_0^w r(s + u) du\right) \bar{F}(s) dG^{(m)}(s)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} \quad (11)$$

When $m \rightarrow \infty$, $\bar{G}^{(m)}(s) \rightarrow 1$ and $G^{(m)}(s) \rightarrow 0$. This means that the failure replacement limit m fails and the proposed warranty is reduced to RFRW. Therefore, the above model can be reduced to a warranty cost $\lim_{m \rightarrow \infty} E[WC] = \int_0^w (c_R - D(x)) dF(x) / \bar{F}(w)$, which is produced by RFRW.

3. The post-warranty random maintenance policies

As mentioned in above, how to model a post-warranty maintenance to sustain the post-warranty reliability has recently received considerable attention. Although the post-warranty maintenance policies to sustain the post-warranty reliability have been investigated extensively, the post-warranty random maintenance policies considering the product working cycle are investigated seldom. In this section, we incorporate the product working cycle into the post-warranty period and investigate the post-warranty random maintenance policies of the product with the warranty proposed in Section 2.

When the product goes through warranty, the total working time of the product is w . This means that reliability is lowered after the product goes through warranty. Therefore, it is necessary to improve reliability of the product through warranty so that the post-warranty period is extended and the post-warranty maintenance cost is reduced. In view of this, we integrate imperfect preventive maintenance (PM) at the warranty period w into the post-warranty maintenance model and investigate two types of the post-warranty random maintenance policies, which will be next provided in Subsection 3.1 and Subsection 3.2.

In order to model conveniently, besides we define similarly the life cycle of the product as an interval from the product installation time to the product replacement occurrence time at the consumer's expense [17, 19], which is composed of the warranty service period and the post-warranty period. By means of this definition, we can derive cost rates model, as below.

3.1. The post-warranty random maintenance policy 1

In this subsection, we introduce imperfect PM at the warranty period w to random periodic replacement last [16] and investigate a post-warranty random maintenance policy satisfying ① imperfect PM is done at the warranty period w ; ② replacement is done at the replacement time T or at the completion of a random working cycle, whichever occurs last; ③ minimal repair removes every failure before replacement. In this article, we refer to this type of maintenance policy as replacement last with PM, which can sustain the post-warranty reliability of the product with random working cycles.

3.1.1. Life cycle cost modeling

In the reliability engineering practice, PM cost is increasing with both the reliability increment resulted from PM and the time where PM is done. The reliability increment resulted from PM is usually estimated by age reduction or/and failure rate reduction [26]. In this article, age reduction is used as a measure of the reliability increment resulted from PM. At the expiry of the proposed warranty, age of the product equates its total working time w . Denote the decreased function $(1 - \varphi(n))w$ with respect to the decision variable n ($n = 0, 1, \dots$) by the reliability increment resulted from PM at w , then PM cost C_{PM} at w can be modeled as an increasing function with both $(1 - \varphi(n))w$ and w , as follows:

$$C_{PM} = c_h ((1 - \varphi(n))w)^{\alpha_2} (w)^{\beta_2} = c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} \quad (12)$$

where c_h is a cost coefficient and satisfies $c_h > 0$; α_2 is an elasticity coefficient of input on reliability improvement and satisfies $\alpha_2 > 0$; β_2 is an elasticity coefficient of input on implementation at w for PM and satisfies $\beta_2 > 0$; $\varphi(n)$ satisfies $0 < \varphi(n) < 1$ where n is the maintenance ability level. Note that when $\varphi(n) = 0$, PM is reduced to perfect PM; when $\varphi(n) = 1$, any maintenance (including PM and minimal repair) is not implemented.

In this article, we have assumed that the random working cycle Y_j ($j = 1, 2, \dots$) is independent identically distributed to the distribution function $G(y)$ with the lack-of-memory property. This assumption means that remaining completion time of a project is still subject to the distribution function $G(y)$. Besides, the probability that replacement is done at the replacement time T or at the completion of a random working cycle, whichever occurs last, can be respectively represented as $G(T)$ and $\bar{G}(T)$. Thus, the costs related to them can be respectively computed as:

$$G(T) \left(c_P - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du \right) \quad \text{and} \\ \int_T^\infty \left(c_P - D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) dG(t), \quad \text{where}$$

$\varphi(n)w$ is virtual age after PM at w ; c_f is unit failure cost resulted from each failure; c_P ($c_P < c_R$) is unit replacement cost suffered for the consumer. By summing, further the expected value $E[C_I(n, T)]$ of the total cost during the post-warranty period is computed as:

$$E[C_I(n, T)] = G(T) \left(c_P - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du \right) + \\ \int_T^\infty \left(c_P - D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) dG(t) + c_P \\ = (c_f + c_m) \int_0^T r(\varphi(n)w + u) du + c_P - D(\varphi(n)w + T) + \int_T^\infty \bar{G}(t) \left(d(\varphi(n)w + t) + (c_f + c_m)r(\varphi(n)w + t) \right) dt \quad (13)$$

where $d(\varphi(n)w+t)$ is first-order derivative with respect to t of the depreciation expense $D(\varphi(n)w+t)$.

By multiplying c_f on the expected number $E[\kappa]$ of failure replacement, the expected value of the total failure cost resulted from all failure replacements can be obtained as $E[\kappa] \cdot c_f = c_f \int_0^w \bar{G}^{(m)}(u) dF(u) / \left(1 - \int_0^w \bar{G}^{(m)}(u) dF(u)\right)$. By replacing c_m in $E[WC_m]$ as c_f , the total failure cost resulted from all failures in the interval $(S_m, w]$ can be obtained, i.e., $c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s) / \left(1 - \int_0^w \bar{G}^{(m)}(u) dF(u)\right)$. Besides, the expected value $E[C_I(n, T)]$ of the total cost during the post-warranty period have been offered in (13) and PM cost C_{PM} at w has been obtained in (12). On the basis of life cycle definition, by summing, the expected value $E[C_I(L)]$ of the life cycle cost is derived as:

$$E[C_I(L)] = \frac{c_f \int_0^w \bar{G}^{(m)}(u) dF(u)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + \frac{c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + C_{PM} + E[C_I(n, T)]$$

$$= \frac{c_f \int_0^w \bar{G}^{(m)}(u) dF(u) + c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} + c_p - D(\varphi(n)w + T) +$$

$$(c_f + c_m) \int_0^T r(\varphi(n)w + u) du + \int_T^\infty \bar{G}(t) (d(\varphi(n)w + t) + (c_f + c_m)r(\varphi(n)w + t)) dt \quad (14)$$

3.1.2. Life cycle length modeling

Until the i^{th} product goes through warranty, the manufacturer performs totally $i-1$ failure replacements. In this case, the total warranty service period resulted from $i-1$ failure replacements can be obtained as $\sum_{k=0}^{i-1} X_k$ where $X_0 = 0$. Since the number $i-1$ of failure replacement satisfies the geometric distribution $p^{i-1}q$, the expected value $E[W]$ of the total warranty service period resulted from all failure replacements can be expressed as:

$$E[W] = E \left[\sum_{i=1}^{\infty} p^{i-1} q \left(\sum_{k=0}^{i-1} X_k \right) \right] = \frac{p}{q} \cdot E[X_k] = \frac{\int_0^w x \bar{G}^{(m)}(x) dF(x)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} \quad (15)$$

where X_k is subject to $H(x)$ in (4) and $E[X_k] = \int_0^w x \bar{G}^{(m)}(x) dF(x) / \int_0^w \bar{G}^{(m)}(u) dF(u)$.

For the product through warranty, its warranty service period is equal to w and the probability that it is replaced at the replacement time T or at the completion of a random working cycle (whichever occurs last) can be respectively computed as $G(T)$ and $\bar{G}(T)$. The corresponding replacement times are respectively $G(T)T$ and $\int_T^\infty u dG(u)$. On the basis of life cycle definition, by summing, the expected value $E[L_I]$ of the life cycle length can be expressed as:

$$E[L_I] = E[W] + w + G(T)T + \int_T^\infty u dG(u) = \frac{\int_0^w x \bar{G}^{(m)}(x) dF(x)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + w + T + \int_T^\infty \bar{G}(u) du \quad (16)$$

3.1.3. Cost rate modeling

The expected value $E[C(L_I)]$ of the life cycle cost and the expected value $E[L_I]$ of the life cycle length have been presented respectively in (14) and (16). Let $A = \left(c_f \int_0^w \bar{G}^{(m)}(u) dF(u) + c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s) \right) / \left(1 - \int_0^w \bar{G}^{(m)}(u) dF(u) \right) + c_p$

and $B = \int_0^w x \bar{G}^{(m)}(x) dF(x) / \left(1 - \int_0^w \bar{G}^{(m)}(u) dF(u) \right) + w$, by the renewal rewarded theorem [1], the expected cost rate $CR_I(n, T)$ can be calculated as:

$$A + c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du +$$

$$CR_I(n, T) = \frac{\int_T^\infty \bar{G}(t) (d(\varphi(n)w + t) + (c_f + c_m)r(\varphi(n)w + t)) dt}{B + T + \int_T^\infty \bar{G}(u) du} \quad (17)$$

Since the expression of $r(u)$ is undefined and unspecific, it is difficult to obtain optimum analytical solutions. But, the existence and uniqueness of optimum solutions can be summarized by discussing the first-order derivative with respect to decision variables of cost rate models. The detail process has been presented and extensively discussed by the literature [19, 22, 31]. In view of this, the existence and uniqueness of optimum solutions are no longer summarized in this article and interested reader consults the above literature, hereinafter similarly.

3.1.4. Special cases

The expected cost rate $CR_I(n, T)$ in (17) is constructed by defining that the proposed warranty is used to ensure reliability preference during the warranty period and by defining that replacement last with PM is used to sustain the post-warranty reliability. By discussing, the expected cost rate $CR_I(n, T)$ can be reduced to some special models representing special problems, as follows:

Case 1: when $m \rightarrow \infty$, model in (17) can be reduced to:

$$A + c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du +$$

$$CR_I(n, T) = \frac{\int_T^\infty \bar{G}(t) (d(\varphi(n)w + t) + (c_f + c_m)r(\varphi(n)w + t)) dt}{B + T + \int_T^\infty \bar{G}(u) du} \quad (18)$$

where $A = F(w)c_f / \bar{F}(w) + c_p$ and $B = \int_0^w \bar{F}(x) dx / F(w)$.

As mentioned in above, $m \rightarrow \infty$ means that the failure replacement limit m is failed and the proposed warranty is reduced to RFRW. Therefore, model in (18) represents an expected cost rate where RFRW is used to warrant the product and replacement last with PM is used to sustain the post-warranty reliability.

Case 2: when $\bar{G}(t) = 0$ and $m \rightarrow \infty$, model in (17) can be reduced to:

$$CR_I(n, T) = \frac{F(w)c_f / \bar{F}(w) + c_p + c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du}{\int_0^w \bar{F}(x) dx / F(w) + T} \quad (19)$$

$\bar{G}(t) = 0$ means that replacement at the completion of a random working cycle is not existed. Therefore, model in (19) represents an expected cost rate where RFRW warrants the product and periodic replacement with PM sustains the post-warranty reliability.

In addition to these special models, some other models are also obtained by discussing one or more of other parameters in the model $CR_I(n, T)$, here we no longer offer them.

3.2. The post-warranty random maintenance policy 2

In this subsection, we introduce imperfect PM at the warranty period w to random periodic replacement first [16] and investigate other post-warranty random maintenance policy satisfying ① imperfect

PM is done at the warranty period w ; ② replacement is done at the replacement time T or at the completion of a random working cycle, whichever occurs first; ③ minimal repair removes every failure before replacement. In this article, we refer to this type of maintenance policy as replacement first with PM to sustain the post-warranty reliability of the product.

Obviously, the unique difference between replacement last with PM and replacement first with PM is that replacement occurrence of the former is decided by 'whichever occurs last' and while replacement occurrence of the latter is decided by 'whichever occurs first'.

3.2.1. Life cycle cost modeling

The probability that replacement is performed at the replacement time T or at the completion of a random working cycle, whichever occurs first, can be respectively represented as $\bar{G}(T)$ and $G(T)$. Besides, the costs related to them can be respectively computed as $\bar{G}(T) \left(c_P - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du \right)$ and $\int_0^T \left(c_P - D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) dG(t)$. By summing, the expected value $E[C_f(n, T)]$ of the total cost during the post-warranty period is computed as:

$$\begin{aligned} E[C_f(n, T)] &= \bar{G}(T) \left(c_P - D(\varphi(n)w + T) + (c_f + c_m) \int_0^T r(\varphi(n)w + u) du \right) + \\ &\quad \int_0^T \left(c_P - D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) dG(t) \\ &= \int_0^T \bar{G}(t) d \left(-D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) + c_P - D(\varphi(n)w) \end{aligned} \quad (20)$$

PM cost C_{PM} at w has been obtained in (12) and the total failure cost resulted from the proposed warranty is $\left(c_f \int_0^w \bar{G}^{(m)}(u) dF(u) + c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s) \right) / \left(1 - \int_0^w \bar{G}^{(m)}(u) dF(u) \right)$, which is similar to (14). On the basis of life cycle definition, by summing, the expected value of the life cycle cost is derived as:

$$\begin{aligned} E[C(L_f)] &= \frac{c_f \int_0^w \bar{G}^{(m)}(u) dF(u) + c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + C_{PM} + E[C_f(n, T)] \\ &= \frac{c_f \int_0^w \bar{G}^{(m)}(u) dF(u) + c_f \int_0^w \left(\int_0^w r(s+u) du \right) \bar{F}(s) dG^{(m)}(s)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} + \\ &\quad \int_0^T \bar{G}(t) d \left(-D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) + c_P - D(\varphi(n)w) \end{aligned} \quad (21)$$

3.2.2. Life cycle length modeling

For the product through warranty, the probability that it is replaced at the replacement time T or at the completion of a random working cycle, whichever occurs first, can be respectively computed as $\bar{G}(T)$ and $G(T)$, and the corresponding replacement times are respectively $\bar{G}(T)T$ and $\int_0^T u dG(u)$. On the basis of life cycle definition, by summing, the expected value $E[L_f]$ of the life cycle length can be expressed as:

$$E[L_f] = E[W] + w + \bar{G}(T)T + \int_0^T u dG(u) = \frac{\int_0^w x \bar{G}^{(m)}(x) dF(x)}{1 - \int_0^w \bar{G}^{(m)}(u) dF(u)} + w + \int_0^T \bar{G}(u) du \quad (22)$$

where $E[W]$ has been offered in (15).

3.2.3. Cost rate modeling

The expected value $E[C(L_f)]$ of the life cycle cost and the expected value $E[L_f]$ of the life cycle length have been offered respectively in (21) and (22). Then, the expected cost rate $CR_f(n, T)$ can be calculated as:

$$CR_f(n, T) = \frac{A + c_h (1 - \varphi(n))^{\alpha_2} (w)^{\alpha_2 + \beta_2} + \int_0^T \bar{G}(t) d \left(-D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) - D(\varphi(n)w)}{B + \int_0^T \bar{G}(u) du} \quad (23)$$

where A and B have been offered in (17).

3.2.4. Special cases

The expected cost rate $CR_f(n, T)$ in (23) is constructed by defining that the proposed warranty warrants the product and by defining that replacement first with PM sustains the post-warranty reliability of the product. The expected cost rate $CR_f(n, T)$ can be reduced to some special models representing special problems, as below.

Case I: when $m \rightarrow \infty$, model in (23) can be reduced to:

$$CR_f(n, T) = \frac{F(w)c_f / \bar{F}(w) + \int_0^T \bar{G}(t) d \left(-D(\varphi(n)w + t) + (c_f + c_m) \int_0^t r(\varphi(n)w + u) du \right) + c_P - D(\varphi(n)w)}{\int_0^w \bar{F}(x) dx / \bar{F}(w) + \int_0^T \bar{G}(u) du} \quad (24)$$

When $m \rightarrow \infty$, $\bar{G}^{(m)}(s) \rightarrow 1$ and $G^{(m)}(s) \rightarrow 0$. Similar to the above discussions, this means that the failure replacement limit m fails and the proposed warranty is reduced to RFRW. Therefore, model in (24) represents an expected cost rate where RFRW is used to warrant the product and replacement first with PM is used to sustain the post-warranty reliability.

Case II: when $\bar{G}(t) = 1$, $n = 0$ and $m \rightarrow \infty$, model in (23) can be reduced to:

$$CR_f(n, T) = \frac{F(w)c_f / \bar{F}(w) + c_P - D(T) + (c_f + c_m) \int_0^T r(w + u) du}{\int_0^w \bar{F}(x) dx / \bar{F}(w) + T} \quad (25)$$

$\bar{G}(t) = 1$ means that replacement at the completion of a random working cycle is removed. $n = 0$ means that PM is not performed and replacement first with PM is reduced to classic periodic replacement policy. Therefore, model in (25) represents an expected cost rate where RFRW warrants the product and classic periodic replacement policy sustains the post-warranty reliability.

Besides, some other models are also offered by discussing one or more of other parameters in the model $CR_f(n, T)$, here we no longer present them.

4. Comparison

Both replacement last with PM and replacement first with PM can sustain the post-warranty reliability of the product. However, making decision on which to sustain the post-warranty reliability is a concerned problem for consumers. In view of this, we present a comparing approach, which can help consumers to make decision on the post-warranty random maintenance policies.

Firstly, let $E[L_f^*]$ and $E[L_f^*]$ be respectively optimum expected values of the life cycle length, which are corresponding to two types of the post-warranty random maintenance policy; secondly, let $E[C(L_f^*)]$ and $E[C(L_f^*)]$ be respectively optimum expected values of the life cycle costs related to two types of the post-warranty random maintenance policy; thirdly, let L_f^{**} and L_f^{**} be respectively cycle lengths related to two types of the post-warranty random maintenance policy, under the case that total costs related to two types of the post-

warranty random maintenance policy are equal. Finally, the comparing approach presented in this article can be summarized as below:

Step 1: Let $L_l^{**} = E[C(L_f^*)] \cdot E[L_l^*]$ and $L_f^{**} = E[C(L_l^*)] \cdot E[L_f^*]$.

Step 2: If $L_l^{**} > L_f^{**}$, then replacement last with PM should be selected to sustain the post-warranty reliability of the product; if $L_f^{**} > L_l^{**}$, then replacement first with PM should be selected to sustain the post-warranty reliability of the product; if $L_l^{**} = L_f^{**}$, then any one of them can sustain the post-warranty reliability of the product because both are equivalent from the performance's perspective.

Note that decision-making result between post-warranty random maintenance policies can also be obtained by comparing total costs related to two types of the post-warranty random maintenance policy, under the case that cycle length of each post-warranty random maintenance policy is equal to a common value. Besides, the comparing approach presented in above can be extended to make decision on three or more the post-warranty random maintenance policies (or maintenance policies).

5. Numerical experiments

The intelligent mobile equipment is frequently used to inspect the remote hidden trouble of the high-voltage electric power equipment. Management can detect operating information of the intelligent mobile equipment by means of the advanced network technology, such as turn on, turn off, failure and working time. The intelligent mobile equipment is powered on when used and is powered off when use is completed. The time interval between power on and power off is a random working cycle.

From the perspective of reliability engineering practice, it is an impossible reality that the product after maintenance is „as good as new”. This means the maintenance ability is limited, namely value n of the maintenance ability level is not infinite. This article uses $\varphi(n) = (n+1)e^{-n}$ to model the reliability alteration resulted from PM, where n ($n = 0, 1, 2, 3, 4, 5$) represents maintenance ability level. The maximum value of maintenance ability level is reached when $n = 5$ and PM is not needed to be performed when $n = 0$.

In order to illustrate the proposed warranty and the policies investigated in this article, assume that lifetime of the intelligent mobile equipment is subject to a two-parameter Weibull function $F(x)$ with a failure rate $r(u) = a(u)^b$, where $a > 0$ and $b > 0$; and assume that working cycle is subject to an exponential distribution function $G(y)$ with a constant failure rate λ (i.e., $G(y) = 1 - \exp(-\lambda y)$) and some constant parameters are offered in Table 1. Other parameters (except decision variables) not to be assigned value in Table 1 are provided when needed.

Table 1. Parameter value

| c_m | c_f | α_1 | β_1 | α_2 | β_2 | c_h | c_R | c_p | a |
|-------|-------|------------|-----------|------------|-----------|-------|-------|-------|-----|
| 0.1 | 0.1 | 0.1 | 1 | 1 | 1 | 0.1 | 10 | 12 | 0.1 |

5.1. Sensitivity analysis of the proposed warranty

In order to illustrate characteristic of the proposed warranty, we plot Figure 1 where $w=2$ and $b=1$. As shown in Figure 1, when the failure replacement limit m increases for a given λ , the warranty cost produced by the proposed warranty increases first and then tends to the warranty cost (i.e., constant warranty cost) produced by RFRW. As mentioned in above, $m \rightarrow \infty$ means that the proposed warranty is transformed into RFRW. Therefore, the above change law with respect to m is existed. This indicates that when the limited number of random working cycle is used as warranty term of the proposed warranty, then the warranty cost produced by the proposed warranty can be reduced compared with traditional RFRW and the manufacturer

can control the warranty cost produced by the proposed warranty by adjusting m . From Figure 1, besides we can find that the warranty cost produced by the proposed warranty is decreasing with respect to λ when the failure replacement limit m is same and is a smaller number.

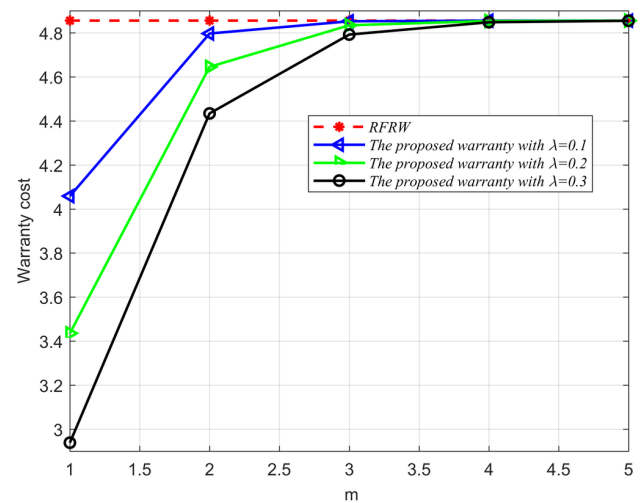


Fig. 1. Warranty cost versus m and λ

In order to further illustrate characteristic of the proposed warranty, we make Figure 2, where $\lambda = 0.1$ and $b = 1$.

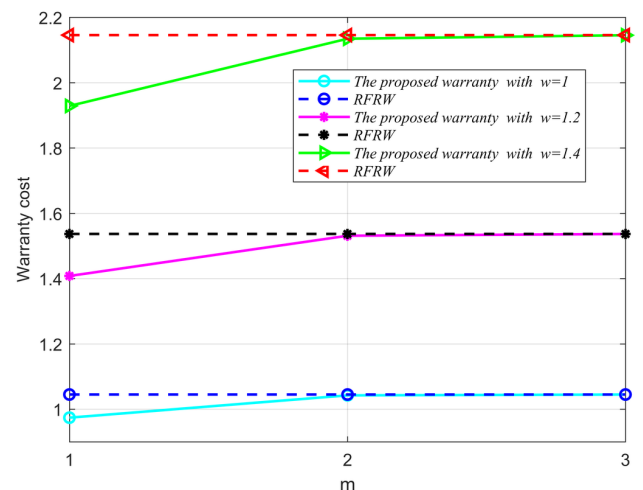


Fig. 2. Warranty cost versus m and w

As shown in Figure 2, the warranty cost produced by the proposed warranty increases first and then tends to a constant warranty cost produced by RFRW as m increases when w is given. This change law indicates that the warranty cost produced by the proposed warranty can be reduced compared with traditional RFRW and the manufacturer can control the warranty cost produced by the proposed warranty by adjusting m , which is similar to conclusions in Figure 1.

5.2. Sensitivity analysis of the post-warranty random maintenance policy 1

For description convenience, in this subsection we represent random periodic replacement first as replacement first, and represent random periodic replacement last as replacement last.

In order to display the existence and uniqueness of the optimum solutions (i.e., n^* and T^*) and the optimum value $CR_l(n^*, T^*)$, we make Figure 3 where $m = 2$, $w = 2$ and $b = 1$.

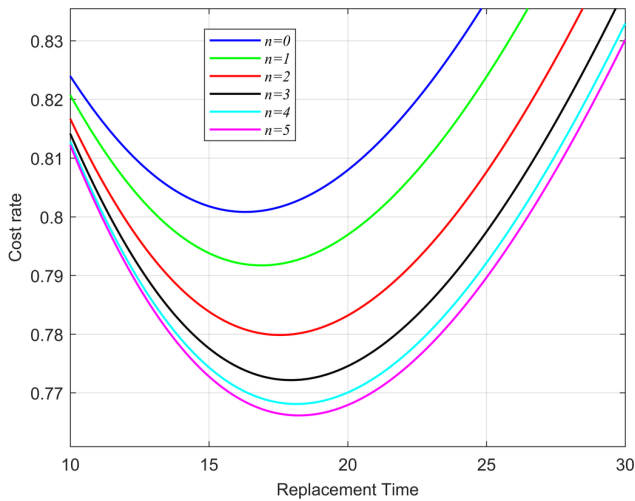


Fig. 3. Optimum solution and optimum value

As shown in Figure 3, the optimum replacement time T^* and the optimum cost rate $CR_f(n^*, T^*)$ are existed uniquely. From Figure 3, we can find that the optimum replacement time T^* is increasing with respect to n , whereas the optimum cost rate $CR_f(n, T^*)$ is decreasing with respect to n . From Figure 3, besides, we can conclude that $n^* = 5$ and replacement last with PM (i.e., $n \neq 0$) is superior to replacement last (i.e., $n = 0$) because replacement last with PM can produce a longer T^* and a lower $CR_f(n^*, T^*)$.

In order to indicate the effect of the failure replacement limit m on replacement last with PM, we make Table 2, where $\lambda = 0.1$, $b = 1$ and $w = 2$.

As shown in Table 2, the optimum replacement time T^* and the optimum cost rate $CR_f(n, T^*)$ decreases gradually to a constant with respect to the failure replacement limit m for a given n . As mentioned in Subsection 3.1.4, the proposed warranty is reduced to RFRW when the failure replacement limit m increases. As m increases, therefore, the optimum replacement time T^* and the optimum cost rate $CR_f(n, T^*)$ decreases gradually to a constant, which is obtained by optimizing the model in (18). From Table 2, besides, we can conclude that $n^* = 5$ and replacement last with PM (i.e., $n \neq 0$) is superior to replacement last (i.e., $n = 0$) for a given m because replacement last with PM can produce a longer T^* and a lower $CR_f(n^*, T^*)$.

5.3. Sensitivity analysis of the post-warranty random maintenance policy 2

In this subsection, we display the existence and uniqueness of the optimum solutions (i.e., n^* and T^*) and the optimum cost rate $CR_f(n^*, T^*)$, and the effect of the failure replacement limit m on replacement first with PM.

Table 2. Sensitivity analysis

| n | $m = 2$ | | $m = 3$ | | $m = 4$ | | $m = 5$ | |
|-----|---------|----------------|---------|----------------|---------|----------------|---------|----------------|
| | T^* | $CR_f(n, T^*)$ | T^* | $CR_f(n, T^*)$ | T^* | $CR_f(n, T^*)$ | T^* | $CR_f(n, T^*)$ |
| 0 | 16.3068 | 0.8008 | 16.2934 | 0.8004 | 16.2926 | 0.8004 | 16.2926 | 0.8004 |
| 1 | 16.8873 | 0.7917 | 16.8747 | 0.7913 | 16.8740 | 0.7913 | 16.8740 | 0.7913 |
| 2 | 17.5560 | 0.7799 | 17.5443 | 0.7795 | 17.5437 | 0.7795 | 17.5437 | 0.7795 |
| 3 | 17.9494 | 0.7722 | 17.9382 | 0.7718 | 17.9375 | 0.7718 | 17.9375 | 0.7718 |
| 4 | 18.1470 | 0.7681 | 18.1361 | 0.7677 | 18.1354 | 0.7677 | 18.1354 | 0.7677 |
| 5 | 18.2395 | 0.7661 | 18.2286 | 0.7658 | 18.2280 | 0.7658 | 18.2280 | 0.7658 |

In order to display the existence and uniqueness of the optimum solutions (i.e., n^* and T^*) and the optimum value $CR_f(n^*, T^*)$, we make Figure 4 where $m = 2$, $w = 2$ and $b = 2$. As shown in Figure 4, the optimum replacement time T^* and the optimum cost rate $CR_f(n^*, T^*)$ are existed uniquely. From Figure 4, we can find that the optimum replacement time T^* is increasing with respect to n , whereas the optimum cost rate $CR_f(n, T^*)$ is decreasing with respect to n . As shown in Figure 4, besides, $n^* = 5$ and replacement first with PM (i.e., $n \neq 0$) is superior to replacement first (i.e., $n = 0$) because replacement first with PM can produce a longer T^* and a lower $CR_f(n^*, T^*)$.

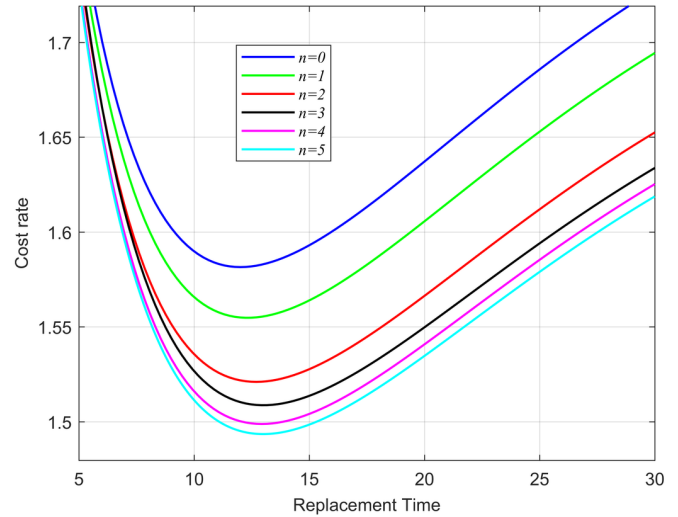


Fig. 4. Optimum solution and optimum value

We make Table 3 where $\lambda = 0.1$, $b = 2$ and $w = 2$. As shown in Table 3, the optimum replacement time T^* and the optimum cost rate $CR_f(n, T^*)$ decreases gradually to a constant with respect to the failure replacement limit m for a given n . The cause of this result is similar to the above analysis. From Table 3, additionally, the optimum replacement time T^* resulted from replacement first with PM (i.e., $n \neq 0$) is greater than the optimum replacement time T^* resulted from replacement first (i.e., $n = 0$), and the optimum cost rate $CR_f(n, T^*)$ resulted from replacement first with PM (i.e., $n \neq 0$) is lower than the optimum cost rate $CR_f(n, T^*)$ resulted from replacement first (i.e., $n = 0$) for a given m . This again means that replacement first with PM is superior to replacement first. From Table 3, thirdly, $n^* = 5$.

5.4. Comparison

Consumers' concern is that which post-warranty random maintenance policy should be used to sustain the post-warranty reliability. This concern is a decision problem. From the consumer's perspective, the post-warranty random maintenance policy with most superior performance is an ideal selection. This indicates that consumers need to make decision on the post-warranty random maintenance policy by comparing performance. In Subsection 5.2 and 5.3, we illustrate performance by comparing optimum replace-

ment time and optimum cost rate. Similarly, here we illustrate performance by comparing optimum replacement time and optimum cost rate.

We make Figure 5 where $\lambda = 0.1$, $m = 2$, $w = 1$, $b = 2$ and $n^* = 5$.

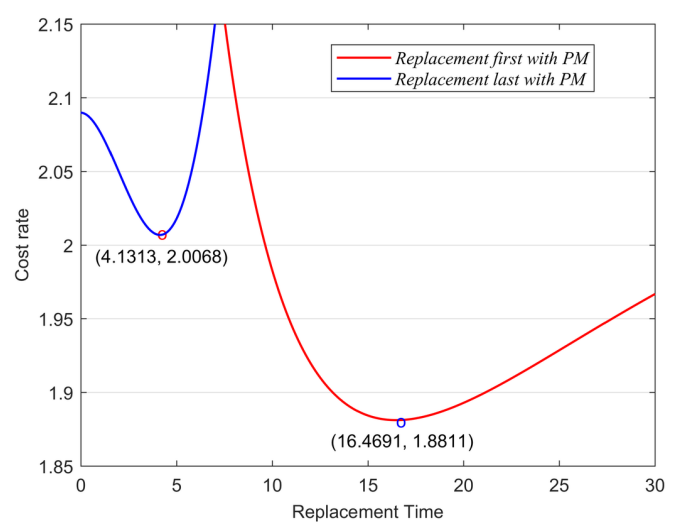


Fig. 5. Comparison

As indicated in Figure 5, optimum cost rate produced by replacement last with PM is greater than optimum cost rate produced by replacement first with PM, whereas optimum replacement time produced by replacement last with PM is not greater than optimum replacement time produced by replacement first with PM. These changes can't rank the post-warranty random maintenance policies because the information provided by them can't manifest any priority order of the post-warranty random maintenance policies.

Next, we use the comparing approach presented in Section 4 to rank the post-warranty random maintenance policies. We make Table 4 where $\lambda = 0.1$, $m = 2$, $b = 2$ and $w = 2$.

Table 4 shows that the cycle length L_l^{**} related to replacement last

Table 4. Comparison

| n | Replacement last with PM | | Replacement first with PM | | Cycle lengths | |
|-----|--------------------------|---------------|---------------------------|---------------|---------------|------------|
| | $E[L_l^*]$ | $E[C(L_l^*)]$ | $E[L_f^*]$ | $E[C(L_f^*)]$ | L_l^{**} | L_f^{**} |
| 0 | 13.7122 | 25.3621 | 10.6994 | 16.9864 | 232.9209 | 271.3593 |
| 1 | 13.8448 | 25.0106 | 10.7892 | 16.8387 | 233.1284 | 269.8444 |
| 2 | 14.0455 | 24.5192 | 10.8965 | 16.6368 | 233.6722 | 267.1735 |
| 3 | 14.1880 | 24.2005 | 10.9613 | 16.5000 | 234.1020 | 265.2689 |
| 4 | 14.2664 | 24.0346 | 10.9942 | 16.4253 | 234.3299 | 264.2412 |
| 5 | 14.3047 | 23.9561 | 11.0097 | 16.3890 | 234.4397 | 263.7495 |

with PM is less than the cycle length L_f^{**} related to replacement first with PM, i.e., $L_l^{**} < L_f^{**}$, when total costs related to two types of the post-warranty random maintenance policy are equal and n is same. This means that the performance of replacement first with PM is superior to the performance of replacement last with PM.

In order to indicate the robustness of the above conclusion, we further make Table 5 where $\lambda = 0.1$, $n^* = 5$, $b = 2$ and $w = 2$. As

shown in Table 5, the cycle length L_l^{**} related to replacement last with PM is lower than the cycle length L_f^{**} related to replacement first with PM, i.e., $L_l^{**} < L_f^{**}$, under the case that total costs related to two types of the post-warranty random maintenance policy are equal and m is same. This indicates again that replacement first with PM is superior to replacement last with PM.

Note that we only analyzed sensitivities of n and m , then we obtained the above conclusion that replacement first with PM is superior to replacement last with PM. If analyzing sensitivities of other parameters, then the conclusion obtained in above may not be established. In either case, the comparing approach presented in Section 4 is a forceful priority method for selection problem of the post-warranty random maintenance policies (or maintenance policies).

6. Conclusions

Taking advanced technologies as the technical background and by designing number of random working cycle as a warranty term, in this article, we proposed a manufacturer's warranty, which can ensure the product reliability performance by monitoring working cycle during the warranty period. The warranty cost produced by the proposed warranty was derived and special model was offered by discussing

warranty term. From the consumer's perspective, we extended the proposed warranty to the post-warranty maintenance and proposed replacement last (and first) with PM, which can sustain the post-warranty reliability by tracking the post-warranty working cycles. Some classic cost rate models representing some special cases were provided by discussing parameters in each cost rate model. We presented a comparing approach to make decision on the post-warranty random maintenance policies. Sensitivities on some key parameters about both the proposed warranty and the proposed post-warranty random maintenance policies were analyzed in numerical experiments. It was discovered that the manufacturer can control the warranty cost when the limited number of random working cycle is used as a warranty term, and it was further discovered that replacement last (first) with PM is more superior compared with replacement last (first).

Acknowledgments

This article is supported by the National Social Science Fund of China (No. 2017BJY008), the Base and Basic Applied Study of Guangdong Province (No. 2020A1515011360), the National Natural Science Foundation of China (No. 71871181).

References

1. Barlow RE, Proschan F. Mathematical theory of reliability. John Wiley & Sons, Hoboken 1965, <http://10.1214/aoms/1177699826>.
2. Cha JH, Finkelstein M, Levitin G. Optimal warranty policy with inspection for heterogeneous, stochastically degrading items. *European Journal of Operational Research* 2021; 289(3): 1142–1152, <https://doi.org/10.1016/j.ejor.2020.07.045>.
3. Chang CC. Optimum preventive maintenance policies for systems subject to random working times, replacement, and minimal repair. *Computers & Industrial Engineering* 2014; 67: 185–194, <https://doi.org/10.1016/j.cie.2013.11.011>.
4. Chen CK, Lo CC, Weng TC. Optimal production run length and warranty period for an imperfect production system under selling price dependent on warranty period. *European Journal of Operational Research* 2017; 259(2): 401–412, <https://doi.org/10.1016/j.ejor.2016.10.038>.
5. Gao H, Cui L, Dong Q. Reliability modeling for a two-phase degradation system with a change point based on a Wiener process. *Reliability Engineering & System Safety* 2020; 193: 106601, <https://doi.org/10.1016/j.ress.2019.106601>.
6. Huang T, Zhao Y, Coit DW, Tang L. Reliability assessment and lifetime prediction of degradation processes considering recoverable shock damages. *IIE Transactions* 2021; 53(5): 614–628, <https://doi.org/10.1080/24725854.2020.1793036>.
7. Hooti F, Ahmadi J, Longobardi M. Optimal extended warranty length with limited number of repairs in the warranty period. *Reliability Engineering & System Safety* 2020; 203: 107111, <https://doi.org/10.1016/j.ress.2020.107111>.
8. Huang HZ, Liu ZJ, Murthy DNP. Optimal reliability, warranty and price for new products. *IIE Transactions* 2007; 39(8): 819–827, <https://doi.org/10.1080/07408170601091907>.
9. He Z, Wang D, He S, Zhang Y, Dai A. Two-dimensional extended warranty strategy including maintenance level and purchase time: A win-win perspective. *Computers & Industrial Engineering* 2020; 141: 106294, <https://doi.org/10.1016/j.cie.2020.106294>.
10. KnopiK L, MigAwA K. Optimal age-replacement policy for non-repairable technical objects with warranty. *Eksploracja i Niezawodność – Maintenance and Reliability* 2017; 19(2): 172–178, <http://dx.doi.org/10.17531/ein.2017.2.4>.
11. Li XY, Chen WB, Li FR, Kang R. Reliability evaluation with limited and censored time-to-failure data based on uncertainty distributions. *Applied Mathematical Modelling* 2021; 94: 403–420, <https://doi.org/10.1016/j.apm.2021.01.029>.
12. Luo M, Wu S. A comprehensive analysis of warranty claims and optimal policies. *European Journal of Operational Research* 2019; 276(1): 144–159, <https://doi.org/10.1016/j.ejor.2018.12.034>.
13. Liu B, Wu J, Xie M. Cost analysis for multi-component system with failure interaction under renewing free-replacement warranty. *European Journal of Operational Research* 2015; 243(3): 874–882, <https://doi.org/10.1016/j.ejor.2015.01.030>.
14. Liu P, Wang G, Su P. Optimal replacement strategies for warranty products with multiple failure modes after warranty expiry. *Computers & Industrial Engineering* 2021; 153: 107040, <https://doi.org/10.1016/j.cie.2020.107040>.
15. Marshall S, Arnold R, Chukova S, Hayakawa Y. Warranty cost analysis: Increasing warranty repair times. *Applied Stochastic Models in Business and Industry* 2018; 34(4): 544–561, <https://doi.org/10.1002/asmb.2323>.
16. Nakagawa T. Random maintenance policies. Springer, London 2014, <https://10.1007/978-1-4471-6575-0>.
17. Park M, Jung KM, Park DH. A generalized age replacement policy for systems under renewing repair-replacement warranty. *IEEE Transactions on Reliability* 2016; 65(2): 604–612, <http://dx.doi.org/10.1109/TR.2015.2500358>.
18. Sánchez-Silva M, Klutke G. Reliability and Life-Cycle Analysis of Deteriorating Systems. Springer, London 2015, <https://doi.org/10.1007%2F978-3-319-20946-3>.
19. Shang L, Si S, Cai Z. Optimal maintenance–replacement policy of products with competing failures after expiry of the warranty. *Computers & Industrial Engineering* 2016; 98: 68–77, <https://doi.org/10.1016/j.cie.2016.05.012>.
20. Shang L, Si S, Sun S, Jin T. Optimal warranty design and post-warranty maintenance for products subject to stochastic degradation. *IIE Transactions* 2018; 50(10): 913–927, <https://doi.org/10.1080/24725854.2018.1448490>.
21. Sheu SH, Liu TH, Zhang ZG, Zhao X, Chien YH. A generalized age-dependent minimal repair with random working times. *Computers & Industrial Engineering* 2021; 156: 107248, <https://doi.org/10.1016/j.cie.2021.107248>.
22. Sheu SH, Liu TH, Zhang ZG. Extended optimal preventive replacement policies with random working cycle. *Reliability Engineering & System Safety* 2019; 188: 398–415, <https://doi.org/10.1016/j.ress.2019.03.036>.
23. Taleizadeh AA, Mokhtarzadeh M. Pricing and two-dimensional warranty policy of multi-products with online and offline channels using a value-at-risk approach. *Computers & Industrial Engineering* 2020; 148: 106674, <https://doi.org/10.1016/j.cie.2020.106674>.
24. Wang L, Pei Z, Zhu H, Liu B. Optimising extended warranty policies following the two-dimensional warranty with repair time threshold. *Eksploracja i Niezawodność - Maintenance and Reliability* 2018; 20(4): 523–530, <http://dx.doi.org/10.17531/ein.2018.4.3>.
25. Wang X, He K, He Z, Li L, Xie M. Cost analysis of a piece-wise renewing free replacement warranty policy. *Computers & Industrial Engineering* 2019; 135: 1047–1062, <https://doi.org/10.1016/j.cie.2019.07.015>.
26. Wu S, Zuo MJ. Linear and Nonlinear Preventive Maintenance Models. *IEEE Transactions on Reliability* 2010; 59(1): 242–249, <http://dx.doi.org/10.1109/TR.2010.2041972>.
27. Ye ZS, Murthy DNP. Warranty menu design for a two-dimensional warranty. *Reliability Engineering & System Safety* 2016; 155: 21–29, <https://doi.org/10.1016/j.ress.2016.05.013>.
28. Yeh RH, Ho WT, Tseng ST. Optimal production run length for products sold with warranty. *European Journal of Operational Research* 2000; 120(3): 575–582, [https://doi.org/10.1016/S0377-2217\(99\)00004-1](https://doi.org/10.1016/S0377-2217(99)00004-1).
29. Zhang S, Zhang Y. Nonlinear mixed reliability model with non-constant shape parameter of aviation cables. *Applied Mathematical Modelling* 2021; 96: 445–455, <https://doi.org/10.1016/j.apm.2021.03.011>.
30. Zhang N, Fouladirad M, Barros A. Evaluation of the warranty cost of a product with type III stochastic dependence between components. *Applied Mathematical Modelling* 2018; 59: 39–53, <https://doi.org/10.1016/j.apm.2018.01.013>.
31. Zhao X, Nakagawa T. Optimization problems of replacement first or last in reliability theory. *European Journal of Operational Research* 2012; 223(1): 141–149, <https://doi.org/10.1016/j.ejor.2012.05.035>.

Influence of the movement of involute profile gears along the off-line of action on the gear tooth position along the line of action direction

Łukasz Jedliński^a

^aLublin University of Technology, ul. Nadbystrzycka 36, 20-618 Lublin, Poland

Indexed by:



Highlights


- Algorithm for finding the exact value of normal backlash is presented.
- Influence of gear parameters on normal backlash for gear OLOA displacement is checked.
- Dynamic simulation is performed and several aspects are analyzed.

Abstract

When gears change their distance along the off-line of action (OLOA) direction, this affects the distance between the working surfaces of the meshing teeth along the line of action (LOA). This effect is usually neglected in studies. To include this effect precise equations are derived for spur gears. The analysis is carried out for the general case of spur gears with shifted profiles frequently used in the industry. The influence of OLOA gear displacement on LOA direction is also a function of gears parameters. An analysis is conducted, and the impact of parameters such as module, pressure angle, gear ratio, and the number of teeth is determined. As an example, a simulation of a 12 DOF analytical model is presented. The movement of gears along OLOA is caused by a frictional force that can be high during tooth degradation e.g. scuffing. Results show that when the movement of gears along the OLOA direction is significant, its influence on the distance between the mating teeth should not be neglected.

Keywords

normal backlash, analytical model, OLOA movement, bearing reactions, dynamic meshing force.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

1. Introduction

A general trend in simulation models for mechanisms and machines is to achieve an accurate solution in acceptable time. Gears and gear units are very often analyzed because of their importance [3, 26]. Analytical models are efficient and based on basic principles ensuring the correctness of results for a given condition [5]. Mechanisms are very often described by equations of motion that do not have an exact solution [18]. In effect, numerical methods must be employed to resolve this problem, and numerical models can be called as semi-analytical. Numerical models are very popular, the dominant numerical technique being finite element analysis (FEA) [16]. FEA models are usually very detailed and can include different types of phenomena. A disadvantage of these models is their very long computation time. To overcome this drawback, hybrid models [12] are used.

All model types are developed to minimize their disadvantages. In this study, analytical formulas are derived to describe spur gears in a more precise and detailed way that can be used in analytical, semi-analytical and hybrid models. A common practice in modeling is to orient the axes of a coordinate system according to the line of action (LOA) and off-line of action (OLOA). This simplifies calculations because then both the normal force and the friction force do not consist of two components. Also, if the center distance is changed along LOA, the same displacement is made between the surfaces of the in-

volute profile teeth that are in mesh. As far as gear movement along OLOA is concerned, its impact on the meshing tooth surface distance is neglected in most studies [4, 6, 17, 21, 24, 27]. In this study, the relationship between the varying distance of gears along OLOA and its effect on the distance between the teeth in mesh is established. Obtained equations are precise for involute gears and no simplifications were made to derive them. To the author's best knowledge, this is a novel solution.

In some studies, the problem of varying center distance and its influence on gear parameters is investigated. A change in the distance between the mating teeth can be considered as varying backlash. In the LOA direction it is normal backlash. In study [13] a single stage spur gear is analyzed. Geometric eccentricity and its influence on backlash are considered. The change in backlash is calculated according to the formula $\Delta b = (r_{b1} + r_{b2})\text{inv}(\alpha) - (r_{b1} + r_{b2})\text{inv}(\alpha')$ where α is the theoretical meshing angle and the theoretical meshing angle α' is 20° . The term "meshing angle" is ambiguous. This formula is also used in [25] and the previous study is quoted as a source. In the study, bearing deformation is the reason for changing the center distance and the influence on backlash is included. Time-varying backlash is defined as $\Delta b = 2(R_1 + R_2)(\text{inv}(\alpha) - \text{inv}(\alpha_0))$ where Δb is the dynamic backlash, α_0 denotes the initial pressure angle for the initial center distance d_0 , $\alpha = \cos^{-1}((R_1 + R_2)/d)$ is the actual pressure

E-mail addresses: Ł. Jedliński - l.jedlinski@pollub.pl

angle. According to these equations, the backlash Δb depends on two parameters: the pressure angle and the center distance. This formula is, however, incorrect. For simplicity, let us consider the movement of one gear. Along the LOA direction, changes in center distance and pressure angle are not significant but have the greatest impact on backlash. In contrast, the movement of a gear along the OLOA direction has a significant impact on center distance and pressure angle, but its impact on backlash is minimal. This stands in contradiction with the above formula, and results will differ by several orders of magnitude, according to the author's calculations. In [7] the effect of eccentricity on backlash is investigated. Time-varying backlash for a driving gear caused by eccentricity is expressed as $\Delta b = -2e_1 \cos(\theta_1 \pm \varphi_1) \tan \alpha$, where e and θ are the gear eccentricity and its initial phase, φ is the angular displacement of gear, and α is the pressure angle. This is a simplified equation. Pressure angle is maintained constant and changes in backlash do not have an exact cosine shape. According to Fig. 1 given in the reference article, the initial phase angle θ is measured from the line which connects the axes of gear rotation. Assuming that θ_1 and φ_1 are equal to 0, the displacement of gears will take place along a direction passing through the axes of rotation. Backlash will reach the maximum value, which is not the case with involute gears. Large bearing clearance and variations in backlash were reported in [14]. The relationship between center distance and backlash was established with approximation of tooth profile. The involute curve was treated as a line. A planetary gear unit used in a turbo-fan engine is studied in [22]. The model presented in the work is comprehensive and many of its parameters are made depended on time. One of them is backlash. The authors derived the formula for calculating backlash from [13], which is incorrect.

From the above paragraph, it can be concluded that information about normal backlash as a function of gear displacement is of vital importance for the model to be more accurate and comprehensive. The relationship between the movement of gears along OLOA and

the distance between the mating tooth surface along LOA (changes in normal backlash) is especially important when the movement (displacement) of gears along OLOA is considerable. Suitable conditions are ensured when e.g. bearing clearance, eccentric gear movement and high friction force are considered in analysis. These three cases will be discussed for different displacement values. An analysis of a rotor-bearing-pedestal system was presented by Cao et al. [1]. With the bearing fit clearance equal to 10 μm and unbalance mass on the rotor, the axis orbit can have a diameter of about 60 μm . The problem of clearance between the rolling bearing outer ring and housing was modelled and analyzed by Chen and Qu [2], who considered a fit clearance of up to 500 μm . Radial internal clearance in the rolling bearing was set to 20 μm in [19, 20]. In a model for analyzing the vibration behavior of a rotor-bearing system, Wang and Zhu [23] set an internal clearance of the cylindrical roller bearing at 60 μm in compliance with ISO 5753-1:2009. Grade accuracy has a great impact on the runout of gears. According to the ISO 1328-2 Cylindrical gears – ISO system of accuracy [9], the runout for gears with a diameter of up to 125 mm can amount to about 200 μm for grade 12. Variations in the center distance reported in [25] amount up to 30 μm and those reported in [22] up to 120 μm . The gear friction coefficient is high when failure occurs. Insufficient lubrication or lack thereof may be its cause. This can lead to scuffing [8, 15]. During this process the friction coefficient has a high value.

For these reasons, it is important to take into account the influence of gear movement along OLOA on distance between the meshing teeth along the LOA direction (normal backlash). An accurate algorithm is derived and an analysis is carried out. In the first simulation the impact of gear parameters such as module, pressure angle, gear ratio and the number of teeth on the distance between the meshing teeth is examined. The other simulation compares results obtained with and without taking into account the movement of gears along OLOA on the dynamic meshing force and bearing force.

Nomenclature

T_m – input motor torque [Nm]

T_d – output device torque [Nm]

I_m – mass moment of inertia of the motor rotor and half of coupling [kg m²]

I_p – mass moment of inertia of the pinion, shaft and half of coupling (pinion subassembly) [kg m²]

I_g – mass moment of inertia of the gear, shaft and half of coupling (gear subassembly) [kg m²]

I_d – mass moment of inertia of the device rotor and half of coupling [kg m²]

I_{px} ($I_{px} = I_{py}$) – mass moment of inertia of the pinion, shaft and half of coupling about y_{op} axis [kg m²]

I_{gx} ($I_{gx} = I_{gy}$) – mass moment of inertia of the gear, shaft and half of coupling about the y_{og} axis [kg m²]

$\ddot{\varphi}$ – angular acceleration [rad/s²]: $\ddot{\varphi}_m$ – motor rotor, $\ddot{\varphi}_p$ – pinion, $\ddot{\varphi}_g$ – gear, $\ddot{\varphi}_d$ – device rotor

$\ddot{\theta}_{px}$ – angular acceleration of the pinion about the y_{op} axis [rad/s²]

$\ddot{\theta}_{gx}$ – angular acceleration of the gear about the y_{og} axis [rad/s²]

$\ddot{\theta}_{py}$ – angular acceleration of the pinion about the x_{op} axis [rad/s²]

$\ddot{\theta}_{gy}$ – angular acceleration of the gear about the x_{og} axis [rad/s²]

\ddot{x}_p – linear acceleration of the pinion on plane defined by the x_{op} axis and pinion axis of rotation [m/s²]

\ddot{x}_g – linear acceleration of the gear on plane defined by the x_{og} axis and gear axis of rotation [m/s²]

x_{py} – distance between new contact point and pinion tooth flank along LOA (x) caused by the movement of gears along OLOA (y) [m],

x_{gy} – distance between new contact point and gear tooth flank along LOA (x) caused by the movement of gears along OLOA (y) [m],

\ddot{y}_p – linear acceleration of the pinion on plane defined by the y_{op} axis and pinion axis of rotation [m/s²]

\ddot{y}_g – linear acceleration of the gear on plane defined by the y_{og} axis and gear axis of rotation [m/s²]

$M_{cm} = c_m(\dot{\varphi}_m - \dot{\varphi}_p)r_m$ – torque applied on the motor coupling from damping [Nm]

$M_{km} = k_m(\varphi_m - \varphi_p)r_m$ – torque applied on the motor coupling from stiffness [Nm]

$M_{cp} = c(r_{b1}\dot{\varphi}_p + \dot{x}_p - \dot{x}_{py} - r_{b2}\dot{\varphi}_g + \dot{x}_g - \dot{x}_{gy})r_{b1}$ – torque applied on the pinion from damping [Nm]

$M_{kp} = k(r_{b1}\varphi_p + x_p - x_{py} - r_{b2}\varphi_g + x_g - x_{gy})r_{b1}$ – torque applied on the pinion from stiffness [Nm]

$M_{kg} = k(r_{b1}\varphi_p + x_p - x_{py} - r_{b2}\varphi_g + x_g - x_{gy})r_{b2}$ – torque applied on the gear from damping [Nm]

$M_{kg} = k(r_{b1}\varphi_p + x_p - x_{py} - r_{b2}\varphi_g + x_g - x_{gy})r_{b2}$ – torque applied on the gear from stiffness [Nm]

$M_{cd} = c_d(\dot{\varphi}_g - \dot{\varphi}_d)r_d$ – torque applied on the device coupling from damping [Nm]

$M_{kd} = k_d(\varphi_g - \varphi_d)r_d$ – torque applied on the device coupling from stiffness [Nm]

$M_{fp} = F_f r_{fp}$ – torque applied on the pinion from tooth friction [Nm]

$M_{fg} = F_f r_{fg}$ – torque applied on the gear from tooth friction [Nm]

$F_n = k(r_{b1}\varphi_p + x_p - x_{py} - r_{b2}\varphi_g + x_g - x_{gy}) + c(r_{b1}\dot{\varphi}_p + \dot{x}_p - \dot{x}_{py} - r_{b2}\dot{\varphi}_g + \dot{x}_g - \dot{x}_{gy})$ – normal force [N]

F_f – tooth friction force [N]

$F_d = \sqrt{F_n^2 + F_f^2}$ – resultant meshing force [N]

r_f – moment arm of sliding friction force [m]

$F_{kblx} = k_{bl}x_{bl}$ – reaction force of bearing 1 from stiffness parallel to the x(LOA) axis [N] (Subscript 2, 3, 4 – bearing 2, bearing 3, bearing 4)

$F_{cblx} = c_{bl}\dot{x}_{bl}$ – reaction force of bearing 1 from damping parallel to the x(LOA) axis [N]

2. Calculation of the dependency between the movement of gears along OLOA on the contact point on LOA

A change in the position of gear axis of rotation along the OLOA direction has impact on the distance between the meshing teeth. From the point of view of dynamic analysis, it is important to know a formula describing changes in the tooth distance along LOA depending on the movement of gears along OLOA. This affects dynamic forces. For clarity of the figure, below is given an example of the displacement of pinion axis of rotation when the center distance is increased without gear movement. Relationships will be derived for a general case describing the displacement of two gears or one gear only.

The nominal position of the gears is marked with a blue dashed line (Fig. 1). The teeth are in mesh at the pitch point C on LOA. According to Fig. 1, the pinion is displaced by a value y_p in the y axis direction. In effect, the distance between the gear axes is increased. To calculate a new center distance a_{w1} , it is convenient to divide the pinion displacement y_p into two separate displacements, e_p and f_p , according to a system of coordinates with the e and f axes:

$$a_{w1} = |O_1O_2| = \sqrt{(a_w + f_p - f_g)^2 + (e_p - e_g)^2} \quad (1)$$

where:

$a_w = |O_1O_2|$ – center distance of gears with shifted profiles,

O_2 – point of intersection with gear axis of rotation,

$f_p = y_p \cos \alpha_w$ – displacement of gear axis of rotation about the f axis,

$f_g = y_g \cos \alpha_w$ – displacement of gear axis of rotation about the f axis,

$e_p = y_p \sin \alpha_w$ – displacement of pinion axis of rotation about the e axis,

$e_g = y_g \sin \alpha_w$ – displacement of gear axis of rotation about the e axis.

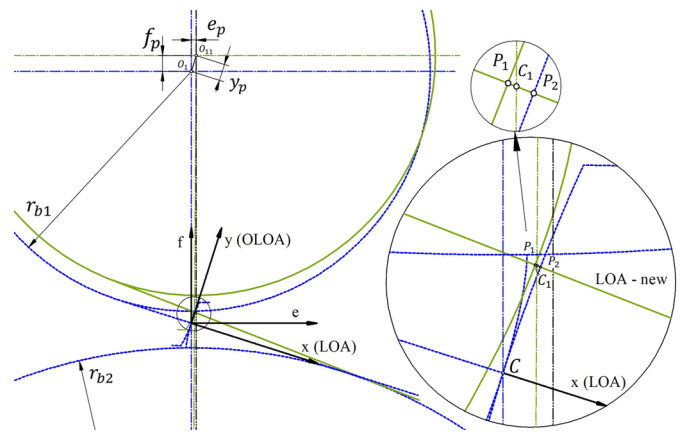


Fig. 1. Pinion displacement by a value y_p about the y axis, leading to a change in pinion axis of rotation position from O_1 to O_{11}

Changes in the center distance of gears have impact on tooth meshing conditions. The green line marks the new LOA. In the magnified image in Fig. 1 one can clearly see a clearance between the mating gear teeth (distance $|P_1P_2|$) at the contact point C before pinion displacement.

For convenience, the distance $|P_1P_2|$ can be divided into two sections. One is the pinion tooth distance P_1 from a new contact point C_1 , while the other is the gear tooth distance P_2 from the contact point C_1 . Fig. 2 shows the pinion along with the relationships enabling the determination of the $|P_1C_1|$ distance. It should be stressed

that the figure is not drawn to scale due to the fact that actual displacements are very small.

The angle ζ between the line connecting the gear center distance before displacement $|O_1O_2|$ and after displacement $|O_{11}O_2|$ is equal to:

$$\zeta = \sin^{-1} \left(\frac{e_p - e_g}{a_{w1}} \right) \quad (2)$$

Next, the angle $\angle U_1O_{11}U_{11}$ denoted by β is calculated as:

$$\beta = \text{inv}\alpha_{w1} - \text{inv}\alpha_w - \zeta \quad (3)$$

where:

$\text{inv}\alpha_{w1} = \tan \alpha_{w1} - \hat{\alpha}_{w1}$ is the involute function,

$\text{inv}\alpha_w = \tan \alpha_w - \hat{\alpha}_w$ is the involute function,

$\hat{\alpha}_{w1} = \cos^{-1} \frac{r_{b1}}{r_{w11}}$ is the contact pressure angle after pinion displacement (for the center distance a_{w1}) [rad],

$\hat{\alpha}_w$ is the contact pressure angle before pinion displacement – nominal position (for the center distance a_w) [rad].

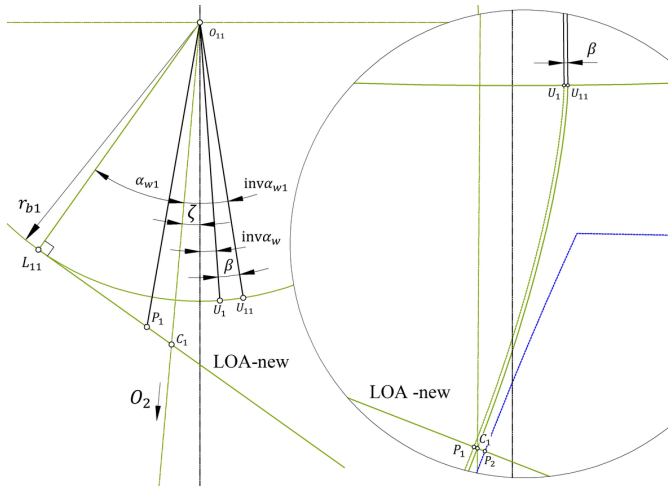


Fig. 2. Relationships for determining the pinion profile distance between P_1 and C_1

The distance $|P_1C_1|$ is equal to the arc length $\widehat{U_1U_{11}}$. This is due to the fact that the involute profile pitch is constant (involute generated on the base circle are spaced by a constant distance measured tangentially to this circle).

$$|P_1C_1| = x_{py} = \hat{\beta} r_{b1} \quad (4)$$

where $\hat{\beta}$ is given in radians.

A similar approach can be adopted to calculate the distance $|P_2C_1|$. By knowing the distance $|P_1C_1|$, the distance between gear tooth profile and new contact point C_1 can be calculated as:

$$|P_2C_1| = x_{gy} = |P_1C_1| \frac{z_2}{z_1} \quad (5)$$

The total displacement resulting from the displacement of gears is:

$$|P_1P_2| = x_{pgy} = |P_1C_1| + |P_2C_1| \quad (6)$$

Formulas (4), (5) and (6) describe simultaneous displacement of two gears, one gear and a pinion. The displacement can be positive or negative about the y axis (OLOA). The gear teeth in contact can take any position on LOA. In all cases, whether the gear axis displacement along OLOA causes a decrease or increase in the center distance, this always results in a clearance between the mating teeth.

The above formulas were derived for a general case in which the gears have shifted profiles. If the gears are without this correction, the following parameters simplify to: $a_w = a$, $\alpha_w = \alpha$, $r_w = r$.

3. Simulation of the influence of gear parameters on the total displacement x_{pgy}

The displacement of gears along the OLOA direction can have different effects on the resultant distance between the mating teeth along LOA x_{pgy} . Tooth size and shape, contact ratio and center distance are the main parameters affecting x_{pgy} , and thus will be investigated in this study. The following properties of gears were selected for simulations, depending on the case under analysis: $m = 3$ mm, $a_w = a = 100$ mm, $z_p = 20$, $z_g = 20$, $\alpha_w = 20^\circ$. In all five cases (Fig. 3-7) only the pinion was displaced ($y_p = -200 \mu\text{m} \div 200 \mu\text{m}$) along OLOA, which affected the nominal center distance.

Figure 3 illustrates the influence of module m . On changing this parameter, the center distance, gear diameters and tooth height change significantly too. The smaller the value of the module is, the greater the total displacement x_{pgy} becomes. This relationship is nonlinear. The maximum value $x_{pgy} = 2.7 \mu\text{m}$ is obtained for $m = 1$ mm and $y_p = -200 \mu\text{m}$ or $200 \mu\text{m}$.

The influence of the gear ratio u is presented in Figure 4. Different values of the gear ratio u are obtained by changing the number of gear teeth $z_g = 16 \div 105$, with the number of pinion teeth maintained constant at $z_p = 20$. By changing the number of gear teeth, the gear diameter, center distance and contact ratio change, too. The smaller value of the gear ratio is, the greater the total displacement x_{pgy} becomes. This relationship is nonlinear. The maximum value $x_{pgy} = 1.1 \mu\text{m}$ is obtained for $u = 0.8$ and $y_p = -200 \mu\text{m}$ or $200 \mu\text{m}$.

The pressure angle α is another investigated parameter. In this case, other parameters do not change like in previous cases, the only exception being the base circle dimension. The smaller the pressure angle value is, the greater the total displacement x_{pgy} becomes. This relationship is nonlinear (Fig. 5). The maximum value $x_{pgy} = 1.8 \mu\text{m}$ is obtained for $\alpha = 11^\circ$ and $y_p = -200 \mu\text{m}$ or $200 \mu\text{m}$.

The last examined parameter is the number of the gear teeth z_p, z_g . Their number is the same ($z_p = z_g = 18 \div 60$) and has impact on the dimension of gears and center distance. To obtain more general results,

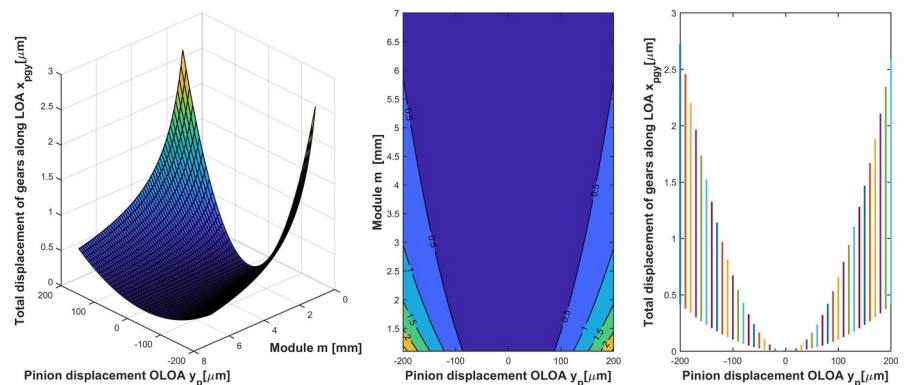


Fig. 3. Relationship between module m , pinion displacement y_p and total tooth displacement x_{pgy} , presented in three types of diagrams

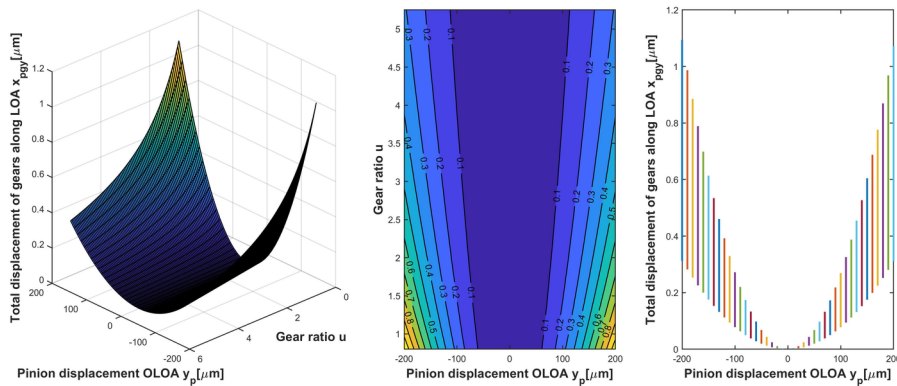


Fig. 4. Relationship between gear ratio u , pinion displacement y_p and total tooth displacement x_{pgy} , presented in three types of diagrams

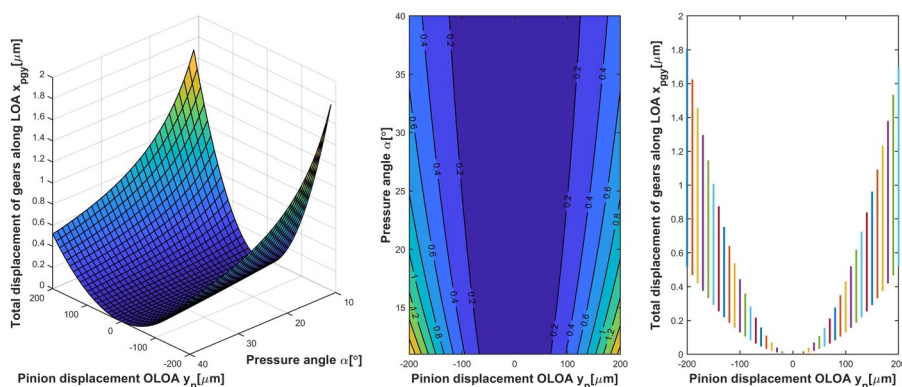


Fig. 5. Relationship between pressure angle α , pinion displacement y_p and total tooth displacement x_{pgy} , presented in three types of diagrams

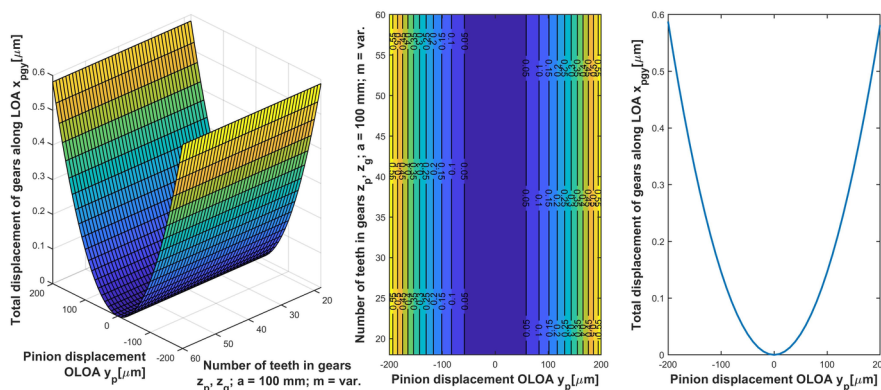


Fig. 6. Relationship between the number of gear teeth z_p, z_g ($a = 100$ mm, $m = \text{var.}$), pinion displacement y_p and total tooth displacement x_{pgy} , presented in three types of diagrams

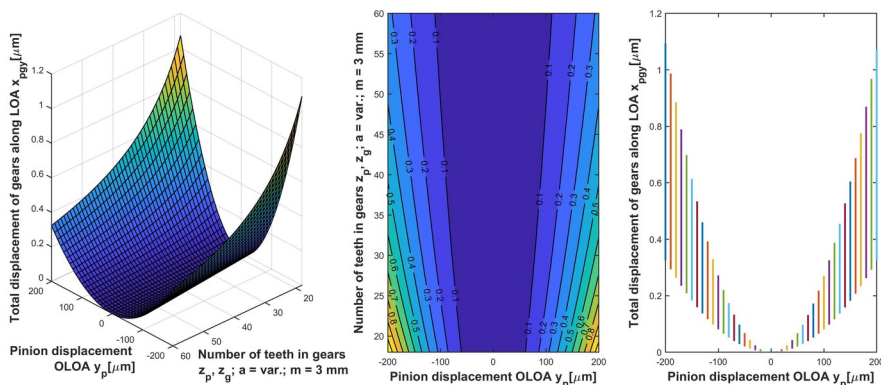


Fig. 7. Relationship between the number of gear teeth z_p, z_g ($a = \text{var.}$, $m = 3$ mm), pinion displacement y_p and total tooth displacement x_{pgy} , presented in three types of diagram

two variants are considered. In the first one (Fig. 6), the center distance a has a constant value of 100 mm, hence the module m must be changed. The total tooth displacement x_{pgy} does not depend on the number of teeth in this case.

In the second variant, the module m is maintained constant at 3 mm, therefore the center distance must vary. Under these conditions, the total tooth displacement x_{pgy} depends on the number of gear teeth. The smaller the number of teeth is, the greater the total displacement x_{pgy} becomes. This relationship is nonlinear (Fig. 7). The maximum value x_{pgy} , 1.1 μm is obtained for $z_p = z_g = 18$ and $y_p = -200$ μm or 200 μm .

It can be seen that the total tooth displacement x_{pgy} is not the same despite the identical pinion displacement y_p along the positive and negative sense of the y axis, which is especially visible in Figure 5. If the pinion displacement along OLOA causes a decrease in the center distance (negative value), its impact on the total tooth displacement is greater. The same relationship can be observed for the gears. It makes no difference whether one gear or two gears move along the OLOA direction. The resulting center distance is a factor affecting the total tooth displacement x_{pgy} . This conclusion can be drawn from Equation (1).

4. Simulation of spur gears for different values of tooth friction coefficient and bearing stiffness

To analyze the influence OLOA displacement of gears on their dynamics, a simulation was performed. One of the situations in which gear displacement along OLOA can be significant is when the force in a radial direction is high. This situation occurs during scuffing. The tooth friction force can achieve significant values as a result of this phenomenon. Nine cases of friction coefficient μ were considered with a step changed every 0.1, from 0.02 to 0.82. Bearing stiffness has a great impact on gear displacement, too. Four values of the bearing stiffness k_b were analyzed: $1.1 \cdot 10^8$ N/m; $1.1 \cdot 10^{8.5}$ N/m; $1.1 \cdot 10^9$ N/m; $1.1 \cdot 10^{9.5}$ N/m. The stiffness values were the same for all bearings and in all directions. Parameters of gear unit and other components are presented in Table 1 and 2.

The simulation was performed on a 12 DOF model (Fig. 8). The model consisted of rigid elements. Every shaft had 5 DOF. The gears were located in the middle of the bearings. The gear unit was connected by couplings with a motor and output device. Dynamic equations were as follows:

$$I_m \ddot{\phi}_m + M_{cm} + M_{km} = T_m \quad (7)$$

$$I_p \ddot{\phi}_p + M_{cp} + M_{kp} = M_{cm} + M_{km} + M_{fp} \quad (8)$$

$$I_g \ddot{\phi}_g + M_{cd} + M_{kd} + M_{fg} = M_{cg} + M_{kg} \quad (9)$$

$$I_d \ddot{\varphi}_d + T_d = M_{cd} + M_{kd} \quad (10)$$

$$F_{kb1x} l_p + F_{cb1x} l_p + m_p \ddot{x}_{pCoM} (l_p - l_{p2}) - I_{px} \ddot{\theta}_{px} = F_n (l_p - l_{p1}) \quad (11)$$

$$F_{kb2x} l_p + F_{cb2x} l_p + m_p \ddot{x}_{pCoM} l_{p2} + I_{px} \ddot{\theta}_{px} = F_n l_{p1} \quad (12)$$

$$F_{kb3x} l_g + F_{cb3x} l_g + m_g \ddot{x}_{gCoM} (l_g - l_{g2}) + I_{gx} \ddot{\theta}_{gx} = F_n (l_g - l_{g1}) \quad (13)$$

$$F_{kb4x} l_g + F_{cb4x} l_g + m_g \ddot{x}_{gCoM} l_{g2} - I_{gx} \ddot{\theta}_{gx} = F_n l_{g1} \quad (14)$$

$$F_{kb1y} l_p + F_{cb1y} l_p + m_p \ddot{y}_{pCoM} (l_p - l_{p2}) - I_{py} \ddot{\theta}_{py} = F_f (l_p - l_{p1}) \quad (15)$$

$$F_{kb2y} l_p + F_{cb2y} l_p + m_p \ddot{y}_{pCoM} l_{p2} + I_{py} \ddot{\theta}_{py} = F_f l_{p1} \quad (16)$$

$$F_{kb3y} l_g + F_{cb3y} l_g + m_g \ddot{y}_{gCoM} (l_g - l_{g2}) + I_{gy} \ddot{\theta}_{gy} = F_f (l_g - l_{g1}) \quad (17)$$

$$F_{kb4y} l_g + F_{cb4y} l_g + m_g \ddot{y}_{gCoM} l_{g2} - I_{gy} \ddot{\theta}_{gy} = F_f l_{g1} \quad (18)$$

Detailed information about tooth stiffness, Coulomb friction and other details concerning the analytical model can be found in [10, 11].

Table 1. Properties of gears

| Parameter | Pinion | Gear |
|--|--|--|
| Number of teeth | $z_p = 20$ | $z_g = 20$ |
| Module [mm] | $m = 2$ | |
| Pressure angle [°] | $\alpha_0 = 20$ | |
| Contact ratio | $\varepsilon = 1,557$ | |
| Moment of inertia (pinion/gear, shaft and half of motor/device coupling) [kgm ²] | $I_p = 0.0033315$; $I_{px} = I_{py} = 0.0117285$ | $I_g = 0.0033315$; $I_{gx} = I_{gy} = 0.0117285$ |
| Mesh damping [Ns/m] | $c = 40$ | |
| Initial angular speed [rad/s] | $\omega_p = 157,0796$ ($n_p = 1500$ rpm) | $\omega_g = 157.0796$ |
| Max stiffness of one pair of teeth [N/m] | $380 \cdot 10^6$ | |

Table 2. Properties of other components

| Parameter | Motor rotor | Device rotor |
|---------------------------------------|---|------------------------|
| Moment of inertia [kgm ²] | $I_m = 0.075$ | $I_d = 0.12$ |
| Torque [Nm] | $T_m = 31.83$ | $T_d = 31.83$ |
| Initial angular speed [rad/s] | $\omega_m = 157.0796$ ($n_m = 1500$ rpm) | $\omega_d = 157.0796$ |
| | Motor coupling | Device coupling |
| Stiffness [N/m] | $k_m = 9.3 \cdot 10^4$ | $k_d = 9.3 \cdot 10^4$ |
| Damping [Ns/m] | $c_m = 10$ | $c_d = 10$ |
| | Bearings | |
| Damping [Ns/m] | $c_b = 40$ | |

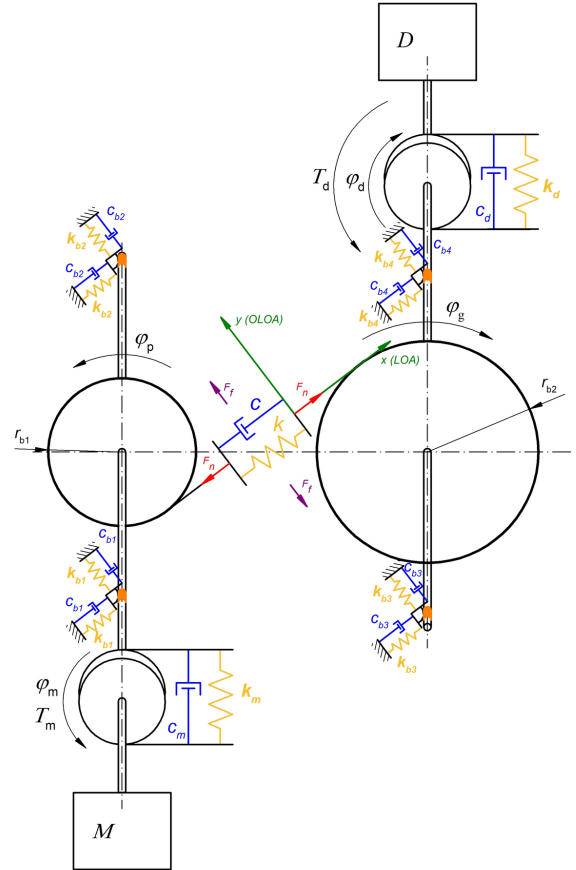


Fig. 8. Analytical model of gear unit with motor M and output device D

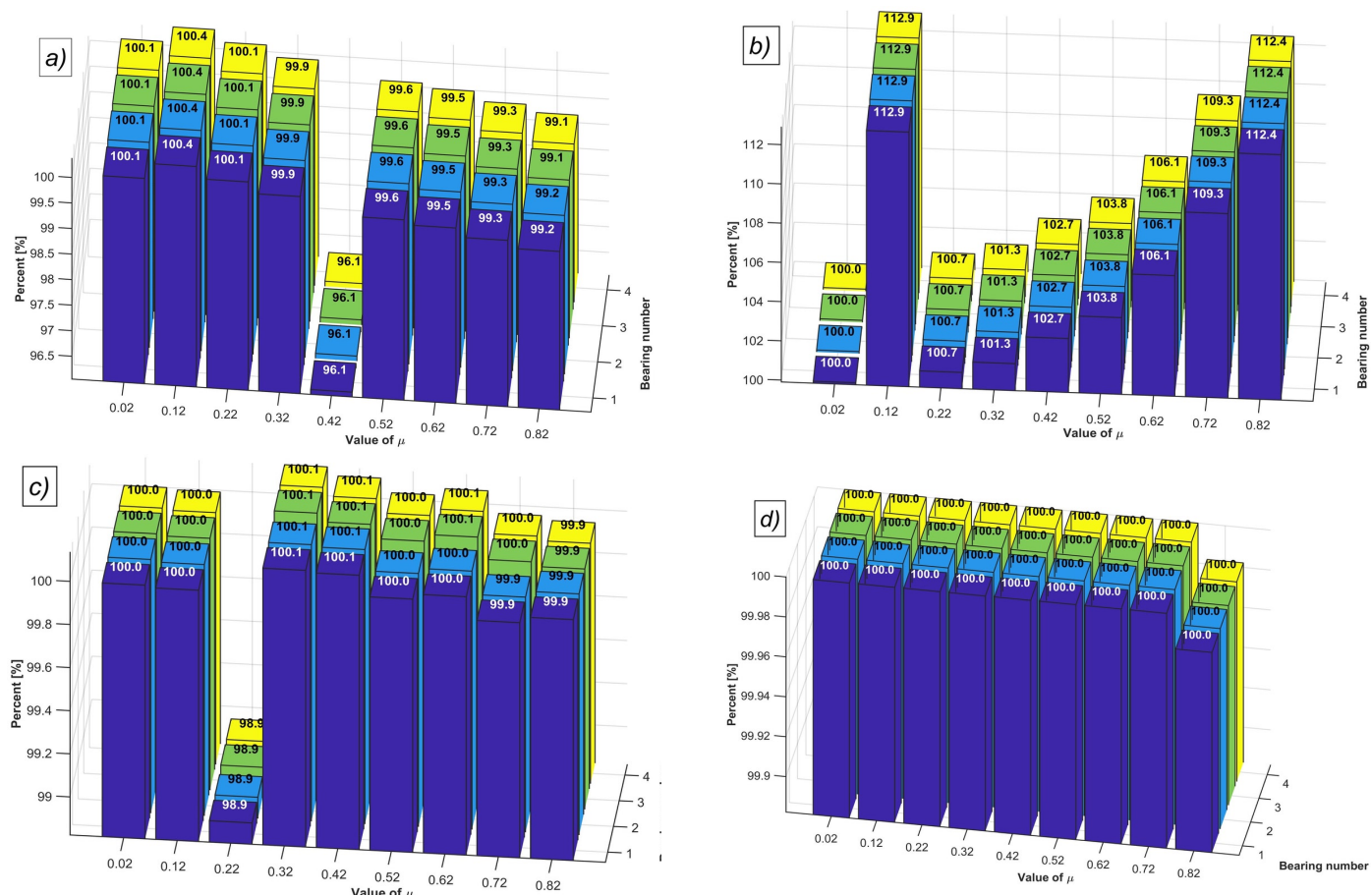


Fig. 9. Percentage change in the resultant reaction force of bearings obtained by considering gear displacement along OLOA and its influence on gear meshing. Results were obtained for the following bearing stiffness values: a) $1,1 \cdot 10^8$ N/m, b) $1,1 \cdot 10^{8.5}$ N/m, c) $1,1 \cdot 10^9$ N/m, d) $1,1 \cdot 10^{9.5}$ N/m

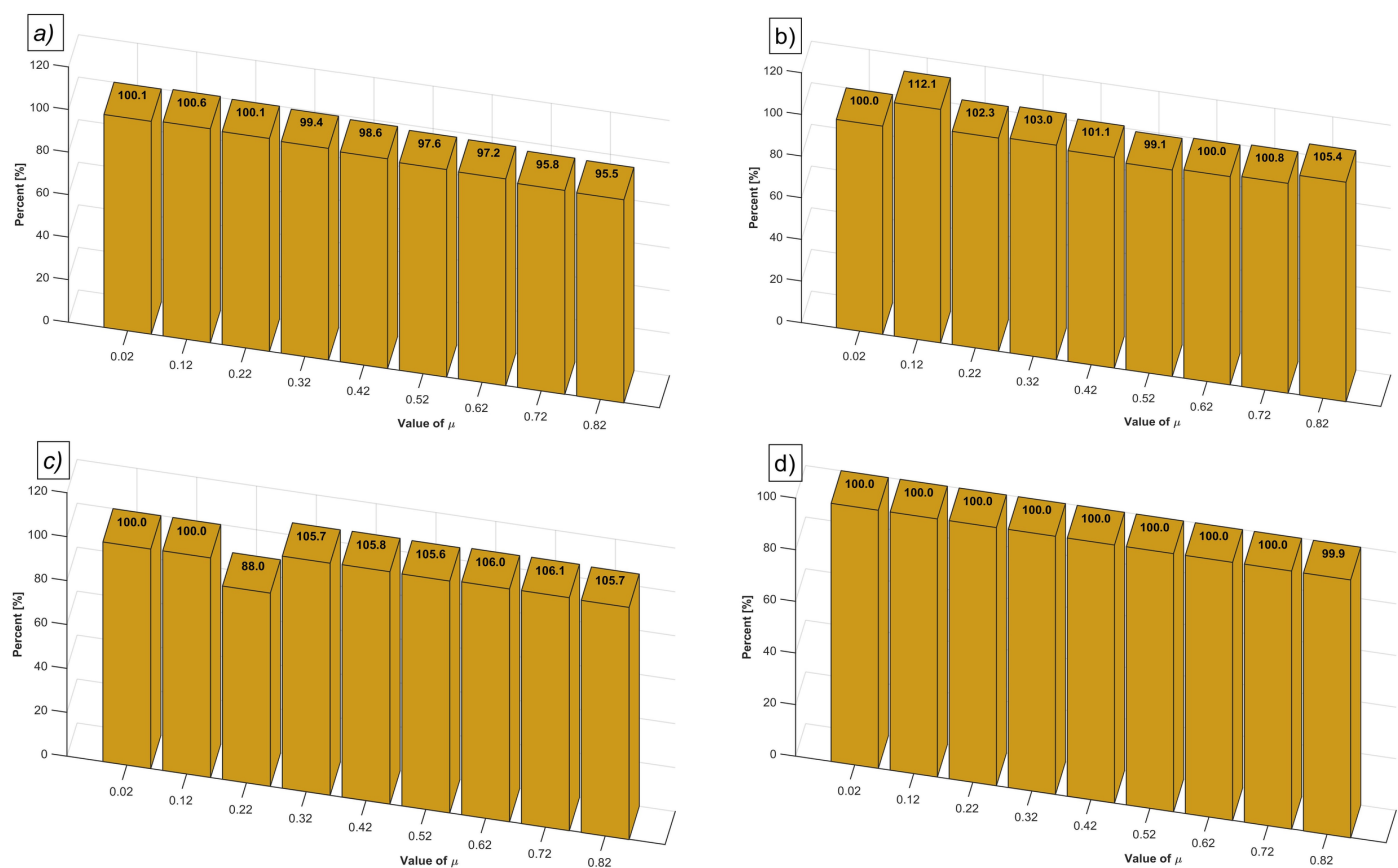


Fig. 10. Percentage change in the resultant meshing force obtained by considering gear displacement along OLOA and its influence on gear meshing. Results were obtained for the following bearing stiffness values: a) $1,1 \cdot 10^8$ N/m, b) $1,1 \cdot 10^{8.5}$ N/m, c) $1,1 \cdot 10^9$ N/m, d) $1,1 \cdot 10^{9.5}$ N/m

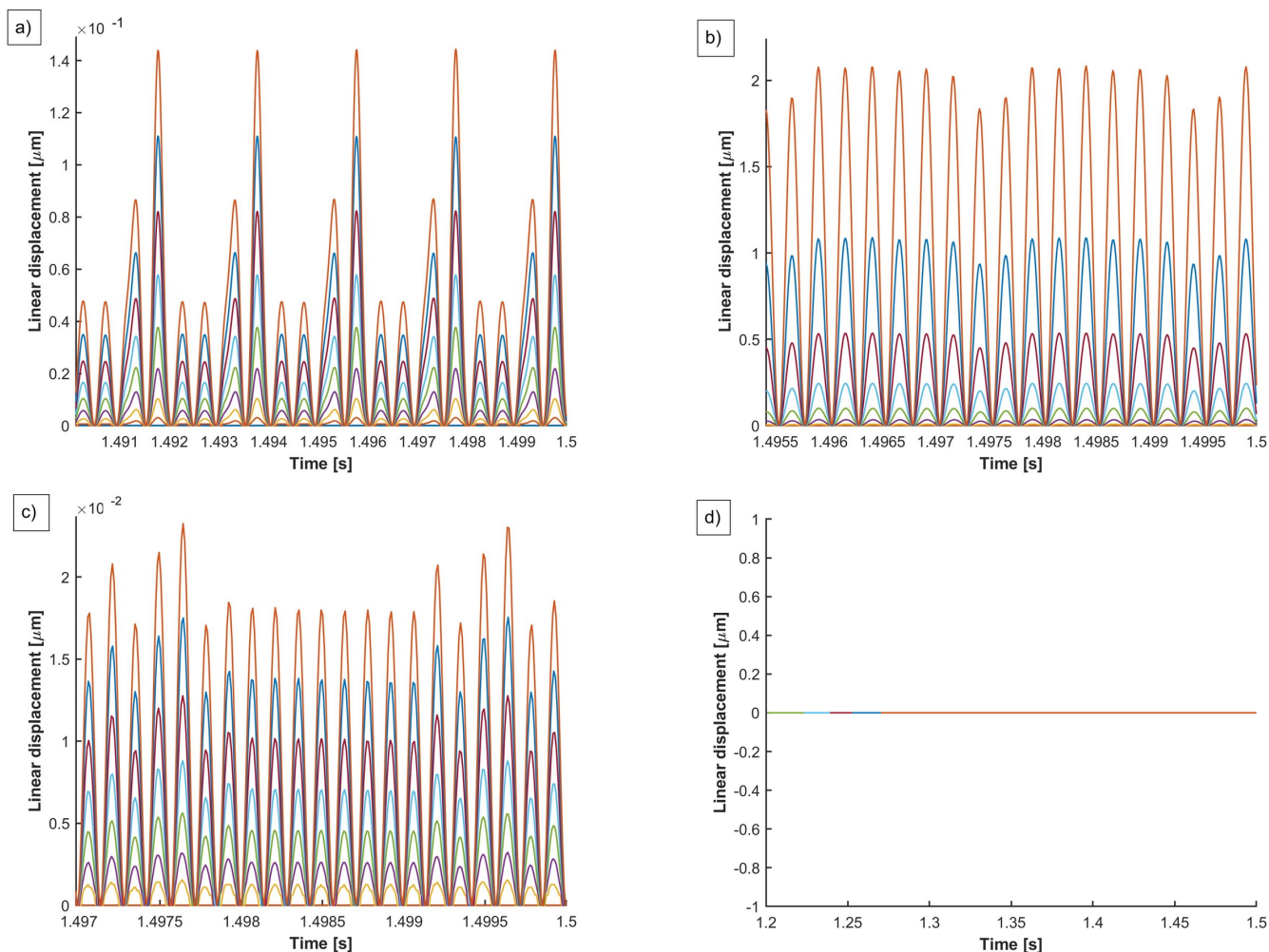


Fig. 11. Total displacement x_{pgy} (increased distance) of the pinion and gear teeth in mesh along the LOA direction as caused by gear displacement along OLOA. Results were obtained for the following bearing stiffness values: a) $1.1 \cdot 10^8$ N/m, b) $1.1 \cdot 10^{8.5}$ N/m, c) $1.1 \cdot 10^9$ N/m, d) $1.1 \cdot 10^{9.5}$ N/m

The reaction force in Fig. 9a is slightly higher than the nominal value for the coefficient of friction ranging from 0.2 to 0.22. For a higher coefficient of friction, the reaction force decreases below the nominal value (nominal value means, that result is obtained without taking into account influence of OLOA displacement of gears on LOA direction). The situation changes when the bearing stiffness is increased (Fig. 9b). For this case, the reaction force is always higher than the nominal value. The difference in the reaction force values in Fig. 9c is very small and in Fig. 9d it is negligible.

The resultant meshing force is a load on the bearings, thus the results in Fig. 9 and Fig. 10 show very similar trends. The maximum difference between the resultant meshing force and the nominal value is 12.1% for the bearing stiffness of $1.1 \cdot 10^{8.5}$ N/m (Fig. 10b). For the bearing stiffness equal to $1.1 \cdot 10^8$ N/m the difference in the resultant meshing is 5.1% (Fig. 10a), while for the bearing stiffness of $1.1 \cdot 10^9$ the difference is 18.1% (Fig. 10c). In Fig. 10d one can only observe one slight change for the highest friction coefficient value.

Examples of waveforms of the total displacement x_{pgy} are presented in Fig. 11. Different colors mark different values of the friction coefficient. The value x_{pgy} is always positive. The maximum displacement is obtained for the bearing stiffness value equal to $1.1 \cdot 10^{8.5}$ N/m. A straight line in Fig. 11d means that displacement does not occur.

5. Conclusions

This study investigated the effect of varying the center distance along OLOA on the gear tooth position along LOA. An exact formula has been derived for a general case of spur gears with shifted profiles. It has been found that changes in the nominal center distance result in

an increased distance between the working surfaces of the gear teeth, i.e. normal backlash. The presented method for determining the distance between the working tooth surface along LOA (normal backlash) is suitable not only for the OLOA direction, but for any other directions, too.

A simulation was performed to establish the relationship between gear parameters and total tooth displacement. It has been found that the module has the greatest impact out of all tested parameters. The second highest result was obtained for the pressure angle. Given that most gear parameters are interdependent, it is not easy to formulate general conclusions. Nonetheless, the movement of gears along the OLOA direction has a greater impact on the movement of the mating teeth along LOA for small gears with a lower gear ratio and a smaller number of teeth.

The effect of the total displacement x_{pgy} on the dynamic behavior of gears was investigated. Based on an analytical model of reaction forces for bearings, resultant meshing force and waveforms of total displacement x_{pgy} were presented. The reaction forces of bearings and the resultant meshing force are strictly interdependent, and the trends obtained for these two parameters are very similar. In the presented example, the reaction forces were higher by more 12 %, and the resultant meshing force was higher, too. The trends are not linear, so a higher frictional force does not always mean that the bearing reaction forces and resultant meshing force will be higher too. It has been found that bearing stiffness has a great impact on the total displacement x_{pgy} and dynamic behavior of gears.

References

1. Cao H, Shi F, Li Y et al. Vibration and stability analysis of rotor-bearing-pedestal system due to clearance fit. *Mechanical Systems and Signal Processing* 2019; 133: 106275, <https://doi.org/10.1016/j.ymssp.2019.106275>.
2. Chen G, Qu M. Modeling and analysis of fit clearance between rolling bearing outer ring and housing. *Journal of Sound and Vibration* 2019; 438: 419-440, <https://doi.org/10.1016/j.jsv.2017.11.004>.
3. Chernets M. Method of calculation of tribotechnical characteristics of the metal-polymer gear, reinforced with glass fiber, taking into account the correction of tooth. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2019; 21(4): 546-552, <https://doi.org/10.17531/ein.2019.4.2>.
4. Cirelli M, Giannini O, Valentini P P, Pennestrì E. Influence of tip relief in spur gears dynamic using multibody models with movable teeth. *Mechanism and Machine Theory* 2020, <https://doi.org/10.1016/j.mechmachtheory.2020.103948>.
5. Dai H, Long X, Chen F, Xun C. An improved analytical model for gear mesh stiffness calculation. *Mechanism and Machine Theory* 2021; 159: 104262, <https://doi.org/10.1016/j.mechmachtheory.2021.104262>.
6. Fernandez-del-Rincon A, Garcia P, Diez-Ibarbia A et al. Enhanced model of gear transmission dynamics for condition monitoring applications: Effects of torque, friction and bearing clearance. *Mechanical Systems and Signal Processing* 2017; 85: 445-467, <https://doi.org/10.1016/j.ymssp.2016.08.031>.
7. Guangjian W, Lin C, Li Y, Shuaidong Z. Research on the dynamic transmission error of a spur gear pair with eccentricities by finite element method. *Mechanism and Machine Theory* 2017; 109: 1-13, <https://doi.org/10.1016/j.mechmachtheory.2016.11.006>.
8. Isaacson A C, Wagner M E. Oil-off characterization method using in-situ friction measurement for gears operating under loss-of-lubrication conditions. *American Gear Manufacturers Association Fall Technical Meeting* 2018: 46-54.
9. ISO 1328-2 1997 - Cylindrical gears-ISO system of accuracy.
10. Jedlinski L. Analysis of the influence of gear tooth friction on dynamic force in a spur gear. *Journal of Physics: Conference Series* 2021, <https://doi.org/10.1088/1742-6596/1736/1/012011>.
11. Jedliński Ł. New Analytical Model of Spur Gears with 5 DOF Shafts and its Comparison with Other DOF Models. *Advances in Science and Technology Research Journal* 2021; 15(1): 79-91, <https://doi.org/10.12913/22998624/130661>.
12. Liu C, Yin X, Liao Y et al. Hybrid dynamic modeling and analysis of the electric vehicle planetary gear system. *Mechanism and Machine Theory* 2020; 150: 103860, <https://doi.org/10.1016/j.mechmachtheory.2020.103860>.
13. Liu H, Zhang C, Xiang C L, Wang C. Tooth profile modification based on lateral- torsional-rocking coupled nonlinear dynamic model of gear system. *Mechanism and Machine Theory* 2016; 105: 606-619, <https://doi.org/10.1016/j.mechmachtheory.2016.07.013>.
14. Liu Z, Liu Z, Zhao J, Zhang G. Study on interactions between tooth backlash and journal bearing clearance nonlinearity in spur gear pair system. *Mechanism and Machine Theory* 2017; 107: 229-245, <https://doi.org/10.1016/j.mechmachtheory.2016.09.024>.
15. Michalczewski R, Kalbarczyk M, Michalak M et al. New Scuffing Test Methods for the Determination of the Scuffing Resistance of Coated Gears. *Tribology - Fundamentals and Advancements* 2013, <https://doi.org/10.5772/54569>.
16. Mohsenzadeh R, Shelesh-Nezhad K, Chakherlou T N. Experimental and finite element analysis on the performance of polyacetal/carbon black nanocomposite gears. *Tribology International* 2021; 160: 107055, <https://doi.org/10.1016/j.triboint.2021.107055>.
17. Shi J fei, Gou X feng, Zhu L yun. Modeling and analysis of a spur gear pair considering multi-state mesh with time-varying parameters and backlash. *Mechanism and Machine Theory* 2019; 134: 582-603, <https://doi.org/10.1016/j.mechmachtheory.2019.01.018>.
18. Skrickij V, Viktor Skrickij Marijonas Bogdevičius Rasa Žygiene. Evaluation of the spur gear condition using extended frequency range. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2017; 19(3): 476-484, <https://doi.org/10.17531/ein.2017.3.19>.
19. Tiwari M, Gupta K, Prakash O. Effect of radial internal clearance of a ball bearing on the dynamics of a balanced horizontal rotor. *Journal of Sound and Vibration* 2000; 238(5): 723-756, <https://doi.org/10.1006/jsvi.1999.3109>.
20. Tomović R. Calculation of the necessary level of external radial load for inner ring support on q rolling elements in a radial bearing with internal radial clearance. *International Journal of Mechanical Sciences* 2012; 60(1): 23-33, <https://doi.org/10.1016/j.ijmecsci.2012.04.002>.
21. Walha L, Fakhfakh T, Haddar M. Nonlinear dynamics of a two-stage gear system with mesh stiffness fluctuation, bearing flexibility and backlash. *Mechanism and Machine Theory* 2009; 44(5): 1058-1069, <https://doi.org/10.1016/j.mechmachtheory.2008.05.008>.
22. Wang S, Zhu R. Theoretical investigation of the improved nonlinear dynamic model for star gearing system in GTF gearbox based on dynamic meshing parameters. *Mechanism and Machine Theory* 2021; 156: 104108, <https://doi.org/10.1016/j.mechmachtheory.2020.104108>.
23. Wang Z, Zhu C. A new model for analyzing the vibration behaviors of rotor-bearing system. *Communications in Nonlinear Science and Numerical Simulation* 2020; 83: 105130, <https://doi.org/10.1016/j.cnsns.2019.105130>.
24. Xiao Y, Fu L, Luo J et al. Nonlinear dynamic characteristic analysis of a coated gear transmission system. *Coatings* 2020; 10(1): 4-6, <https://doi.org/10.3390/coatings10010039>.
25. Yi Y, Huang K, Xiong Y, Sang M. Nonlinear dynamic modelling and analysis for a spur gear system with time-varying pressure angle and gear backlash. *Mechanical Systems and Signal Processing* 2019; 132: 18-34, <https://doi.org/10.1016/j.ymssp.2019.06.013>.
26. Zhang X, Zhao J. Compound fault detection in gearbox based on time synchronous resample and adaptive variational mode decomposition. *Eksplatacja i Niezawodność - Maintenance and Reliability* 2020; 22(1): 161-169, <https://doi.org/10.17531/ein.2020.1.19>.
27. Zhao Z, Han H, Wang P et al. An improved model for meshing characteristics analysis of spur gears considering fractal surface contact and friction. *Mechanism and Machine Theory* 2021; 158: 104219, <https://doi.org/10.1016/j.mechmachtheory.2020.104219>.

Remaining useful life prediction with insufficient degradation data based on deep learning approach

Yi Lyu^{a,*}, Yijie Jiang^b, Qichen Zhang^c, Ci Chen^b

^a University of Electronic Science and Technology of China Zhongshan Institute, School of Computer, Zhongshan, China, 528400

^b Guangdong University of Technology, Guangzhou, School of Automation, China, 510006

^c University of Electronic Science and Technology of China, School of Computer Science and Engineering, Chengdu 611731

Indexed by:



Highlights


- Focus on the improvement of the RUL prediction effect in the case of insufficient degradation data.
- A data amplification network based on cycleGAN is designed to effectively increase the size of the degradation dataset.
- A RUL prediction framework is constructed with the sliding time window strategy and BiLSTM network.
- Experimental results show the RUL prediction performance has been significantly improved by the proposed data amplification approach.

Abstract

Remaining useful life (RUL) prediction plays a crucial role in decision-making in condition-based maintenance for preventing catastrophic field failure. For degradation-failed products, the data of performance deterioration process are the key for lifetime estimation. Deep learning has been proved to have excellent performance in RUL prediction given that the degradation data are sufficiently large. However, in some applications, the degradation data are insufficient, under which how to improve the prediction accuracy is yet a challenging problem. To tackle such a challenge, we propose a novel deep learning-based RUL prediction framework by amplifying the degradation dataset. Specifically, we leverage the cycle-consistent generative adversarial network to generate the synthetic data, based on which the original degradation dataset is amplified so that the data characteristics hidden in the sample space could be captured. Moreover, the sliding time window strategy and deep bidirectional long short-term memory network are employed to complete the RUL prediction framework. We show the effectiveness of the proposed method by running it on the turbine engine data set from the National Aeronautics and Space Administration. The comparative experiments show that our method outperforms a case without the use of the synthetically generated data.

Keywords

deep learning, remaining useful life, degradation data, data amplification, cycle-consistent generative adversarial network.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

1. Introduction

Accurate prediction of the remaining useful life (RUL) is extremely valuable for decision-making in condition-based maintenance for preventing catastrophic field failure. For degradation-failed products, the data of performance deterioration process plays a major role in RUL estimating. The methods of RUL estimation can be divided into three categories: 1) method based on failure mechanism analysis [9, 22], 2) method based on data-driven approach, and 3) hybrid method that combines the first two. The key point of RUL prediction using the first method is to fully understand the degradation mechanism of the target equipment. Prior knowledge in the target field is indispensable when establishing a mathematical model of the degradation process. However, as the complexity of the equipment increases and automation advances, obtaining complete knowledge of the degradation mechanism becomes difficult [7, 14]. The aircraft turbine engine data set of the National Aeronautics and Space Administration (NASA) was built from more than ten sensors. These data should be analyzed together to reveal the health indicators of the turbine engine. Different

from the method based on failure mechanism analysis, the data-driven approach does not require researchers to have a comprehensive understanding of the target equipment [12, 23]. After collecting sufficient degradation data from sensors, researchers could construct a nonlinear mapping between degradation data and the real equipment health indicators, and meanwhile solves the dynamic dependency problems [8, 28]. This nonlinear mapping network can be used to predict the RUL of the equipment used on site.

Data-driven methods, especially the deep learning approach have developed substantially in recent years [3, 6, 16–18, 24]. Considering the problem of weak dependence of time-series information, Zhu [36] combined the information of the previous convolutional layer with the current layer and proposed a multiscale convolutional neural network (CNN) for RUL prediction. The long-range dependence problem exists in many studies on time-series data. Li [11] selected the long short-term memory network (LSTM) and CNN as the base model to build the RUL prediction model. LSTM can save past information for the current network parameter update and CNN has a

(*) Corresponding author.

E-mail addresses: Y. Lyu - lvyi913@zsc.edu.cn, Y. Jiang - yijiejiang@live.com, Q. Zhang - zhangqichen0708@163.com, C. Chen - gduccc@gmail.com

strong ability in local feature extraction. The combination of the two improves the accuracy of the prediction network. Group method of data handling-type neural network (GMDH) can self-organize and generate the optimal network structure based on the training data [22]. Ge [4] generates three GMDH networks through different division of training data, and integrates the results of the three GMDH networks with a three-layer back propagation (BP) neural network to solve the disadvantage of local optimum of GMDH and improve the generalization ability. A.Ragab [19] developed a data-driven prognostic methodology using both the age and condition monitoring data as inputs, which can deal with any number of condition indicators. Under different test conditions, different workloads, environmental conditions and noise levels may lead to different distribution of training set and test set. To solve this problem, Wen [26] used domain-adversarial neural network (DANN) and proposed a data-driven framework with domain adaptability using a bidirectional gated recurrent unit (BGRU). This method can effectively reduce the impact on the performance of RUL prediction due to the different distribution of training data and testing data. Deep learning methods are adopted to address the RUL prediction issue of a specific field, such as bearings [20, 36], lithium-ion batteries [32, 34], lathe tool wear [37], and nuclear systems [38].

Nevertheless, the estimation effect of these mentioned methods is highly dependent on the capacity of the degradation data set. That means the scale of the dataset available in model training phase has a great influence on the RUL prediction accuracy [13]. Abdulraheem [1] explored the effect of the dataset size on prediction results under supervised learning techniques, their findings showed that the model with the largest dataset had the best prediction effect under three datasets listed as dataset size of 400, 800, and 1200. The larger the dataset is, the better is the model established. However, in many actual industrial production practices, obtaining a largescale dataset is not realistic due to the longer degradation time and high cost of collecting degradation data. The XJTUSY rolling bearings dataset mentioned by Wang [25] only collected the complete life cycle of 15 bearings (type LDK UER204), the entire life cycle is only 42h and 18min. Many restrictions on obtaining large-scale degradation data restrict the further development of deep learning data-driven methods in RUL prediction. Moreover, for those newly emerging equipment, there is also a lack of degradation data. Under these scenarios, the RUL prediction performance will be severely affected. Hence, how to improve the prediction accuracy with insufficient degradation data is yet a challenging task.

In the case of insufficient degradation data, the low accuracy of RUL prediction is mainly caused by the low sample diversity, which can be effectively improved by data augmentation [29]. Generative adversarial network (GAN) is a common data augmentation strategy, which can capture the characteristics hidden in the sample space and enrich the diversity of samples [30]. Yoon [31] applied the GAN to the task of generating medical data and produced a patient electronic health dataset containing discrete time series data. In the sequence data generation task, Li [10] utilized GAN to capture the temporal correlation of time series distributions, the generator and discriminator inside the GAN adopt the LSTM network as the basic network, which is friendly to time-series data. Subsequently, Xie [27] generated bearing datasets for various working conditions based on the cycle-consistent generative adversarial network (CycleGAN) framework and its GAN discriminator was trained for fault diagnosis.

Based on the above research, this study developed a complete framework to improve the RUL prediction performance when degradation data is insufficient. Four steps are involved in this framework. Firstly, constructing a data amplification model using the LSTM network which is also as the Generator inside the CycleGAN and mining the inherent distribution of existing degradation data samples of a machine. Second, a data preprocessing strategy is designed for time-series degradation data before they are sent to the augmentation network. Third, the obtained amplified data are preprocessed using sliding time window method and their labels for prediction model training

are obtained. Finally, a data-driven method is built with amplified data for RUL prediction. The contributions of this study are summarized as follows:

- Proposed an amplification network for generating time series degradation data based on CycleGAN; this method uses a small amount of data to train CycleGAN and uses the designed generator based on the LSTM network for data amplification without excessive prior knowledge of the data.
- Designed a data preprocessing strategy to resize the time-series degradation data before they are sent to the designed amplification network.
- Constructed a data-driven RUL prediction model and integrated the above work into a complete set of RUL prediction methods, which is suitable for the degradation data of time-series.
- Compared the performance differences between RUL prediction models trained with amplified data obtained from various amounts of degradation data.

The rest of this paper is organized as follows. Theoretical foundation of the CycleGAN is introduced in section II. Proposed an amplification network based on LSTM and related theory of data preprocessing strategy and RUL prediction model constructed are introduced in section III. An experiment is introduced in section IV. The conclusions are summarized in section V.

2. Theoretical Foundation

CycleGAN is a type of unsupervised learning generative network that was designed to solve the problem of image-to-image translation in the field of vision and graphics by learning the mapping between a set of aligned image pairs from source domain to target domain. The key to achieve this function is an adversarial structure composed of two networks called generator and discriminator. The generator captures the distribution of the true image and constructs a fake one, and the discriminator estimates the probability that the image came from the true image rather than the generator. Ideally, the discriminator's recognition success rate should be approximately equal to 0.5, which means that the discriminator cannot distinguish whether the test image is real or generated, that is, the generator obtained the true mapping between image pairs. To ensure improved learning efficiency, we built a cycle-consistent structure from two directions. Two generators and two discriminators are used in each direction; one of the generators is used to transform the data from *domainA* to *domainA*, and the other generator aims to reconstruct the generated data back to *domainA*. The structure is shown in Figure 1.

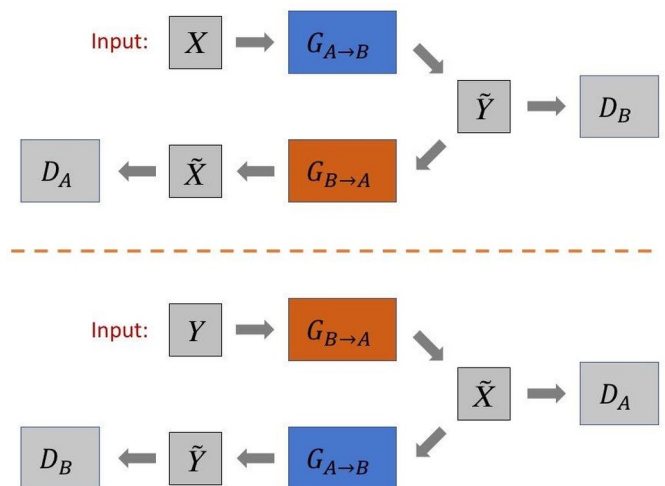


Fig. 1. Structure of CycleGAN

There two types of data X with *domainA* and data Y with *domainB*. In the upper part, data $X = \{x_A^1, x_A^2, \dots, x_A^m\}$ from *domainA*

are sent into the generator $G_{A \rightarrow B}$ randomly. The generated data $\tilde{Y} = \{\tilde{y}_B^1, \tilde{y}_B^2, \dots, \tilde{y}_B^m\}$ obtained with probability distribution are similar to $domainB$, Discriminator D_B distinguishes the generated data \tilde{Y} and from the real data $Y = \{y_B^1, y_B^2, \dots, y_B^m\}$. The generated data \tilde{Y} obtained through $G_{A \rightarrow B}$ are sent to generator $G_{B \rightarrow A}$. The reconstructed data $\tilde{X} = \{\tilde{x}_A^1, \tilde{x}_A^2, \dots, \tilde{x}_A^m\}$ obtained from the generator $G_{B \rightarrow A}$ are distinguished with the real data X of $domainA$ via discriminator D_A .

Value function is shown in Formula 1. To simplify the function, we define $G_{A \rightarrow B}$ as G and $G_{B \rightarrow A}$ as F .

$$\begin{aligned} \min_G \max_{D_Y} \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ = \mathbb{E}_{y \sim p_{data}(y)} [\log D_Y(y)] \\ + \mathbb{E}_{x \sim p_{data}(x)} [\log(1 - D_Y(G(x)))] \end{aligned} \quad (1)$$

In the process of optimizing this value function, the distribution of the data generated by the generator G is updated close to $domainB$, and the discriminator D_Y distinguishes the generated data from the real data. The value function aims to minimize the generation error of G , and maximize the recognition success rate of D_Y . Similarly, we can obtain the value function of another generator F and discriminator D_X :

$$\begin{aligned} \min_F \max_{D_X} \mathcal{L}_{GAN}(F, D_X, Y, X) \\ = \mathbb{E}_{x \sim p_{data}(x)} [\log D_X(x)] \\ + \mathbb{E}_{y \sim p_{data}(y)} [\log(1 - D_X(F(y)))] \end{aligned} \quad (2)$$

Combining both two parts shown above can obtain a cycle-consistency loss:

$$\begin{aligned} \mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)} [\|F(G(x)) - x\|_1] \\ + \mathbb{E}_{y \sim p_{data}(y)} [\|G(F(y)) - y\|_1] \end{aligned} \quad (3)$$

The value function is shown as Formula 4:

$$\begin{aligned} \operatorname{argmin}_{G, F} \max_{D_Y, D_X} \mathcal{L}_{GAN}(G, F, D_X, D_Y) \\ = \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ + \mathcal{L}_{cyc}(G, F) \end{aligned} \quad (4)$$

3. Methodology

In the task of RUL prediction with data-driven approach, the actual effect of the model is largely determined by the data size. Insufficient run-to-failure degradation data are the key to limit the reliability of the prediction model. This work focuses on how to mine potential data distribution information from limited samples and improve the effect of the RUL prediction model using deep learning technology.

In the model our primary hypothesis is that the time series degradation data used to construct RUL predictions are scarce. If the deep learning method is directly used to summarize the degradation features from the limited degradation data and perform RUL prediction, then the prediction effect will not be as good as expected. We proposed a method that consists of three parts. The first part is an amplification network designed by the LSTM network, which can mine the data distribution information from known samples to expanding sample

size [15]. The second is a designed data preprocessing strategy. Owing to the time-dependent dynamic characteristics of the degradation data, the sliding time window strategy is used to fix the dynamic degradation information of the data and adjust the size of the degradation data before sending them to the amplification network to improve the network processing efficiency. The third is described as follows: using the amplified data obtained from the first part to construct a prediction network mainly based on bidirectional long short-term memory (BiLSTM); in the training process, the cyclic neural structure in BiLSTM can effectively solve the problem of long-range dependence in time series, obtain the optimal parameters of the model through the backpropagation algorithm, and construct an RUL prediction network to predict the samples.

3.1. Data Amplification Network Based on CycleGAN

In CycleGAN, using the data of two different domains, the generator can make the mutual conversion of the data from the two domains through the adversarial with the discriminator. To obtain the information of the sparse degradation data in our hypothesis, we replaced the data of the two domains with the degradation data of a single domain. Unlike the previous CycleGAN in which the two generators learned the distribution information from one domain, the scheme we proposed aims to learn from each other with scarce degradation data, and the trained generator is used to complete the generation of degradation data.

The generator based on the LSTM was designed as the amplification network. LSTM is a type of recurrent neural network whose structure contains units with functions such as forgetting and remembering; this network is suitable for processing time series data [35]. In actual situations, the degradation data of the device is usually strongly correlated with time and can be used to solve long-range dependence problems [21].

To establish a connection in the calculation unit cycle at each moment, three gate structures in LSTM was designed, namely, forget gate layer, input gate layer and output gate layer. These gate structures control the information flow at different times, and store short-term time-step dependent information for network parameter update, which alleviate the problem of gradient disappearance or gradient explosion of the classic neural network structure during backpropagation. The LSTM cell structure at time t is shown in Figure 2. The input of the current moment consists of the data from current moment input and the data from previous output, the input of the next moment is composed of the data from the current moment output and the data from the next moment input. The related formula is shown as follows:

Forget gate layer:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

Input gate layer:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (8)$$

Output gate layer:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t * \tanh(C_t) \quad (10)$$

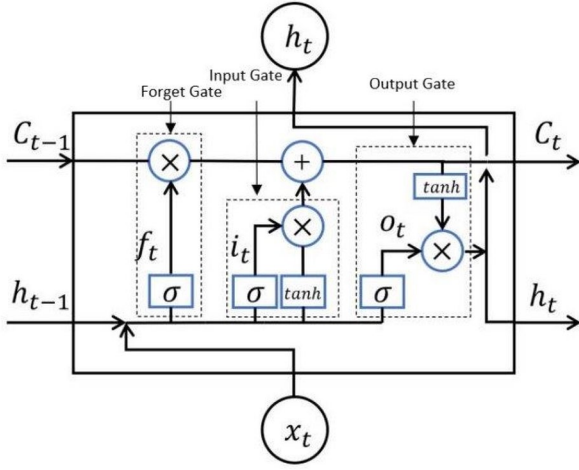


Fig. 2. Cell structure of LSTM at time t

where σ is the activation function, W_f, W_i, W_C, W_o are the weight matrices, x_t is the input data at time t , h_t is the output data at time t , C_t represents the information flow participates in parameters updated throughout the entire training process.

As the degradation data is basically a continuous time series, we improved the output form of the LSTM network and fixed the input and output sizes of the generator network to be consistent to improve the spatial structure of the sequence to reduce the loss of degraded information. Specifically, the dimensions of the input and output should be consistent. We saved the output obtained from each h_t of LSTM from timestep 1 to timestep m , which are used to form the final output from the network. The dimension of the output could be a series instead of a scale. The series can meet the requirements of the network for the input data with time dynamic characteristics. The schematic is shown in Figure 3. On the left is an input data with dimension $n \times s$, where n represents the length of input data and s represents the dimension of sensors in input data. In the center is the generator with timesteps equal to m . On the right is the first output data with dimension $m \times s$. The second output data are obtained with a dense operation at dimension $n \times s$.

The dimension of the input data $n \times s$ is given by the task, where n represents the length of input data and s represents the dimension of sensors. The timestep of LSTM is about to set a larger number than the length of the input data. In Figure 3, timestep ts is set to m , where $m > n$. In the training process, the first line of the input data $1 \times s$ is sent to the generator, the output of the generator with size of $1 \times \text{timesteps}$ consists of values obtained from each timestep. After all the input data are sent into the network, all the outputs are combined into a matrix of dimension $n \times m$. Finally, a dense operation is performed to obtain an output data with size consistent the input data.

To ensure that the generated data are similar to the real data in distribution and avoid the difference of actual generated data that affects the characterization of degradation information, we add maximum mean difference (MMD) into the generator's loss function, which is shown as follows:

$$J(G) = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2} y_i - \tilde{y}_i^2 \right) + MMD \quad (11)$$

where $J(G)$ is the loss function of the generator, n is the number of samples, y_i is the generated sample of i -th instance, and \tilde{y}_i is the target sample of i -th instance.

MMD was designed to measure the difference in data distribution by comparing the statistical information of the two sets of data and was used as a training objective functions for generating networks. In

practice, the inner product between the two samples is replaced with the kernel calculation, and the MMD formula is as follows:

$$\begin{aligned} MMD = & \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n K(x_i, x_j) \\ & - \frac{2}{mn} \sum_{i=1}^n \sum_{j=1}^m K(x_i, y_j) \\ & + \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m K(y_i, y_j) \end{aligned} \quad (12)$$

The inner products are replaced by Gaussian kernel between two samples, and the formula is as follows:

$$K(x, y) = \exp\left(-\|x - y\|^2 / (2\sigma^2)\right) \quad (13)$$

where σ is the bandwidth. We select a group of different σ , and the calculated MMD is averaged as the final value.

In the training process, we optimize the parameters of the generated model by gradient descent algorithm. The samples generated by the model further reduce the difference between the target samples and enable them to meet the task requirements.

3.2. Data Preprocessing Strategy for Amplification

The RUL of the degradation data for training under ideal conditions should be clear. However, even the same type of equipment has a various life cycle due to different qualities or operating environments. To accurately characterize the temporal dynamics of degradation data, we need a data preprocessing strategy before the degradation data with different life cycles is sent to the amplification network.

The strategy of processing data with inconsistent length of life span is as follows. We obtained the initial value of the rapid data degradation stage through statistical analysis. The initial value of the rapid degradation stage divides the degradation data into a normal stage and a rapid degradation stage. We retain the values of the rapid degradation stage. Then, the process of resizing the data occurs in the normal stage, because the value in the normal stage usually maintains a small range of changes and the significance of predicting the RUL in the normal stage is not as important in the rapid degradation stage.

Given time-series degradation data $X_{s,n}$ with size $s \times n$, as shown in Formula 14, we obtain the output $X_{s,n'}$ that meets the requirements with size $s \times n'$.

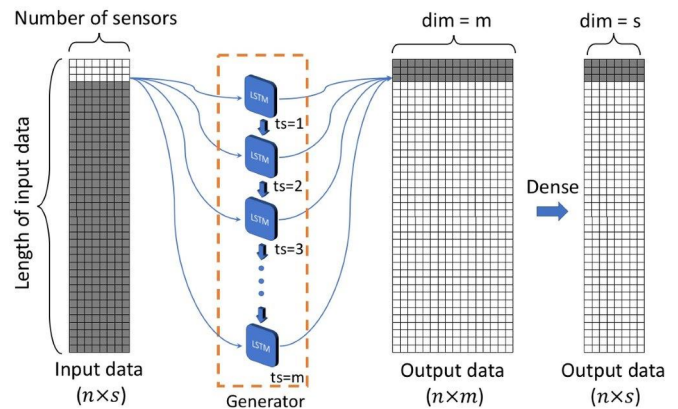


Fig. 3. Structure of the generator based on LSTM

$$\begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{s,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{s,2} \\ \vdots & \vdots & & \vdots \\ x_{1,m} & x_{2,m} & \cdots & x_{s,m} \\ \vdots & \vdots & & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{s,n} \end{bmatrix} \Rightarrow \{X\}_{s \times n} \quad (14)$$

where s represents the number of data features and n represents the life span of the degradation data. We assume that the initial value of the rapid degradation stage obtained by the statistical analysis is m .

In the rapid degradation stage, the value is directly retained without any processing. In the normal stage, two types of resize data strategies are proposed as follows:

- 1) If the current degradation data length is more than n' , then we remove the excess part directly to obtain the data that meets the requirements as follows:

$$\begin{bmatrix} x_{1,l} & x_{2,l} & \cdots & x_{s,l} \\ x_{1,l+1} & x_{2,l+1} & \cdots & x_{s,l+1} \\ \vdots & \vdots & & \vdots \\ x_{1,m} & x_{2,m} & \cdots & x_{s,m} \\ \vdots & \vdots & & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{s,n} \end{bmatrix} \Rightarrow \{X'\}_{s \times n'} \quad (15)$$

The size of the processed data is $s \times n'$, where $l = n - n'$. The excess part is removed from the beginning.

- 2) If the length of the current degradation data is shorter than n' , then we design a data padding strategy. We calculate the average value of the same sensor data in the first time window as the padding data. The substituted x' value for sensor s is expressed as Formula 16.

$$x'_s = \frac{\sum_{n=1}^{L_{tw}} (x_{s,n})}{L_{tw}} + \frac{\gamma}{2} \quad (16)$$

where L_{tw} is the length of time window, and γ is a Gaussian noise in the range of $(x_{s,n}^{min} - x_{s,n}^{max})$, which are the maximum and minimum values of the data at sensor s in one time window. The processed data are shown as follows:

$$\begin{bmatrix} x'_{1,1} & x'_{2,1} & \cdots & x'_{s,1} \\ x'_{1,2} & x'_{2,2} & \cdots & x'_{s,2} \\ \vdots & \vdots & & \vdots \\ x'_{1,l} & x'_{2,l} & \cdots & x'_{s,l} \\ x_{1,1} & x_{2,1} & \cdots & x_{s,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{s,2} \\ \vdots & \vdots & & \vdots \\ x_{1,m} & x_{2,m} & \cdots & x_{s,m} \\ \vdots & \vdots & & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{s,n} \end{bmatrix} \Rightarrow \{X'\}_{s \times n'} \quad (17)$$

where $n^0 = l + n$.

3.3. Data Degradation Strategy

The degradation data of the generated network should be processed into the same dimensions as the data during training. The time-series degradation data can be expressed as follows:

$$\begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{s,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{s,2} \\ \vdots & \vdots & & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{s,n} \end{bmatrix} \Rightarrow \{X\}_{s \times n} \quad (18)$$

where s is the number of data features; for instance, bearings data may have features such as vibration, rotation speed, and temperature. n represents the length of the data on the time scale, which can reflect the working time or service life of the data; this value is directly related to RUL.

All the real degradation data are sent into the CycleGAN for training the generator. The first batch of degradation data are sent into the trained generator to obtain the first batch of amplified data. The degradation data of the next batch is obtained from the amplified data of the previous batch, and the amplification is stopped until a predetermined amount of amplified data is obtained. To ensure that the amplified data retains more original degradation information during the iterative process, the number of iterative amplifications should not be excessive.

3.4. RUL Prediction Model Construction

- 1) *Sliding Time Window Strategy*: For RUL prediction on time-series degradation data, the problem of label identification needs to be solved. One of the intuitive and efficient methods is the sliding time window method [11, 15, 33].

For example, given data sample $X = (x_1, x_2, \dots, x_n)$, $n = 1, 2, 3, \dots, n$, where n is the length of the data sample on the time scale. we specify the sliding time window size l , then k time windows are obtained which $k = \frac{n}{l} + 1$. Each time window can be expressed as follows and the schematic is shown in Figure 4.

$$\begin{aligned} X^1 &= (x_1, x_2, \dots, x_l) \\ X^2 &= (x_{l+1}, x_{l+2}, \dots, x_{2l}) \\ &\vdots \\ X^{k-1} &= (x_{(k-1)l+1}, x_{(k-1)l+2}, \dots, x_{(k-1)2l}) \\ X^k &= (x_{n-l}, \dots, x_{n-1}, x_n) \end{aligned}$$

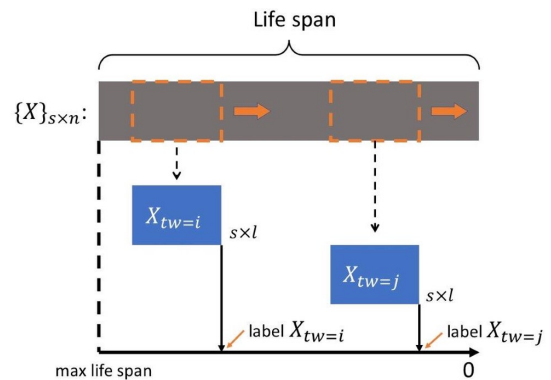


Fig. 4. Schematic of Sliding Time Window

where $X_{tw=i}$ is the i th window. The time window records a piece of information of the degradation data. For complete degradation data, we can obtain k pieces of degradation data and the RUL label of each segment in order.

- 2) *Prediction Model*: A non-linear mapping from data to labels is built by a data-driven method with sufficient labeled data. We use the deep BiLSTM network [5] to build a prediction model. The difference between LSTM and BiLSTM is that the latter increases the reverse transmission process of data information and contains more hidden layers. The structure of BiLSTM is shown in Figure 5.

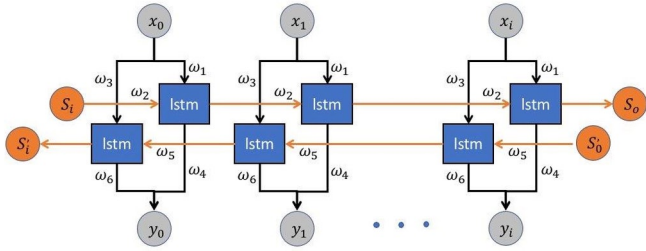


Fig. 5. Structure of BiLSTM

The final output y_i of the bidirectional LSTM consists of three parts: input of the model, input of the forward propagation process, and input of the reverse propagation process:

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \quad (19)$$

$$h'_t = f(w_3 x_t + w_5 h'_{t+1}) \quad (20)$$

$$y_t = g(w_4 h_t + w_6 h'_t) \quad (21)$$

where w_{1-6} represents network parameters, x_t is the input in timestep t , h_t is the value from the forward propagation process, h'_t is the value from reverse propagation process, and g is the activation function.

Owing to the flexibility and versatility of the BiLSTM, a deep network with a stronger non-linear fitting ability was obtained, which is beneficial for RUL prediction by stacking the BiLSTM into three layers. Under this framework, the architecture of a mapping between time window and RUL tag is established, as presented in Figure 6.

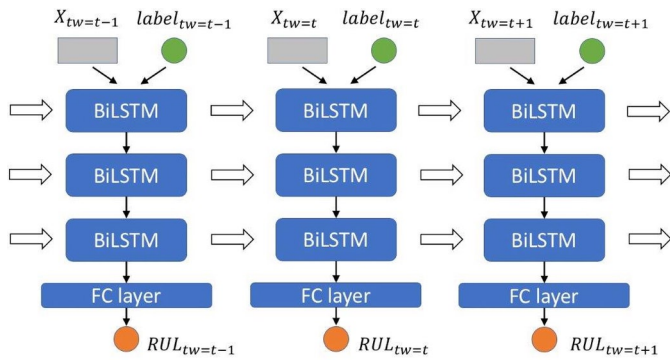


Fig. 6. Structure of RUL prediction model

The main components of the framework are composed of two parts. The first part is a deep learning network composed of stacked BiLSTM. The deep architecture has strong representation capabilities and can learn the time dynamic characteristics between time window degradation data. The other part is a fully connected neural network for regression tasks. Data from stacked BiLSTM which contains degradation information, are used to obtain the predicted RUL from the activation function with $\text{ReLU}(f(x) = \max(0, x))$.

- 3) *RUL prediction objective*: The parameters in the prediction network are obtained through the back propagation through time (BPTT) algorithm and the given value function is shown as Formula 22. It's defined as the error between the model output and the label:

$$L_{rul}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 \quad (22)$$

where $\theta = [w_{1-6}]$ is parameter set of the prediction model and n is the number of units in one batch, y_i and \tilde{y}_i are the model output and label of i -th instance respectively.

3.5. Algorithm Summary

Algorithm of data amplification and RUL prediction is summarized in Algorithm 1. The entire flowchart of data amplification and RUL prediction is shown in Figure 7.

Algorithm 1 Algorithm of data amplification and RUL prediction

Input: Historical degradation data $X_{s,n}$

Data Amplification :

- 1: Process the original training data by data preprocessing strategy to meet the requirements of the amplification network.
- 2: Train the CycleGAN network to obtain a generator based on designed LSTM.
- 3: Use the amplification data as the input to the generator and repeat this step until the specified amount of training data is obtained.

Output: Amplified Data

RUL prediction :

- 4: Use the sliding time window method to obtain training data and corresponding labels.
- 5: Build a RUL prediction network using BiLSTM.
- 6: Predict the RUL of with the prediction network.
- 7: **return** RUL

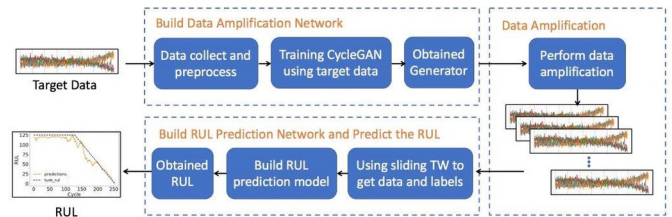


Fig. 7. Flowchart of data amplification and RUL prediction

4. Experiment

An experiment was conducted to validate that our proposed data amplification strategy can improve the prediction effect by data-driven methods when using insufficient training data. We selected the degradation data with the multi-sensor turbo aero engine dataset from NASA. This dataset contains the operational data of the complete life cycle of multiple turbo aero engines, and each engine contains multiple sensor data. The multi-sensor degradation data have higher requirements for the RUL prediction model and show the universality of our proposed methods.

4.1. Data Preprocessing and Analysis

The turbo-aero engine dataset is divided into four sub-datasets: FD001, FD002, FD003 and FD004. Differences only exist in operating conditions and failure modes, and no dependency exists among the sub-datasets. In this experiment, FD001 was selected as the experimental dataset. FD001 contains the complete degradation data of 100 turbine aero engines. the maximum life span is 362, which means that the entire working cycle of this turbine aero engine is 362. Details of the dataset are shown in Table I.

Table I. C-MAPSS dataset

| items | FD001 | FD002 | FD003 | FD004 |
|---------------------------|-------|-------|-------|-------|
| Engines in dataset | 100 | 260 | 100 | 248 |
| Conditions | 1 | 6 | 1 | 6 |
| Fault Modes | 1 | 1 | 2 | 2 |
| Maximum life span(cycles) | 362 | 378 | 525 | 543 |
| Minimum life span(cycles) | 128 | 128 | 145 | 128 |

The sensors are located in all important parts of the turbine aero engine and record the possible parameters related to corresponding degradation indicators. Data from more sensors are considered to provide comprehensive information on engine degradation. Details are shown in Table II.

Table II. C-MAPSS sensors dataset

| Num | Symbol | Description | Units | trend |
|-----|-----------|---------------------------------|------------|-------|
| 1 | T2 | Total temperature at fan inlet | °R | ~ |
| 2 | T24 | Total temperature at LPC outlet | °R | ↑ |
| 3 | T30 | Total temperature at HPC outlet | °R | ↑ |
| 4 | T50 | Total temperature at LPT outlet | °R | ↑ |
| 5 | P2 | Pressure at fan inlet | psia | ~ |
| 6 | P15 | Total pressure in bypass-duct | psia | ~ |
| 7 | P30 | Total pressure at HPC outlet | psia | ↓ |
| 8 | Nf | Physical fan speed | rpm | ↑ |
| 9 | Nc | Physical core speed | rpm | ↑ |
| 10 | epr | Engine pressure ratio (P50/P2) | - | ~ |
| 11 | Ps30 | Static pressure at HPC outlet | psia | ↑ |
| 12 | phi | Ratio of fuel flow to Ps30 | pps psi | ↓ |
| 13 | NRf | Corrected fan speed | rpm | ↑ |
| 14 | NRc | Corrected core speed | rpm | ↓ |
| 15 | BPR | Bypass Ratio | - | ↑ |
| 16 | farB | Burner fuel-air ratio | - | ~ |
| 17 | htBleed | Bleed Enthalpy | - | ↑ |
| 18 | Nf dmd | Demanded fan speed | rpm | ~ |
| 19 | PCNfR dmd | Demanded corrected fan speed | rpm | ~ |
| 20 | W31 | HPT coolant bleed | lbm/s | ↓ |
| 21 | W32 | LPT coolant bleed | lbm/s | ↓ |

A total of 21 sensors were used. Among them, 14 were related to the potential degradation mechanism during the entire degradation process; these sensors are numbered 2, 3, 4, 7, 8, 9, 11, 12, 13, 14, 15, 17, 20, and 21. In the stage of data preprocessing, to avoid any interference of useless information, we select the information of these 14 sensors as target data. Most of the equipment can be divided into normal stage and rapid degradation stage in its life cycle. For the purpose of

RUL prediction, the prediction of RUL in the stage of rapid degradation is more important than the that under normal stage. According to the work of Babu [2], when 125 cycles remain, a clear degradation trend appears. The degradation failure threshold is set to 125 cycles, as shown in Figure 8.

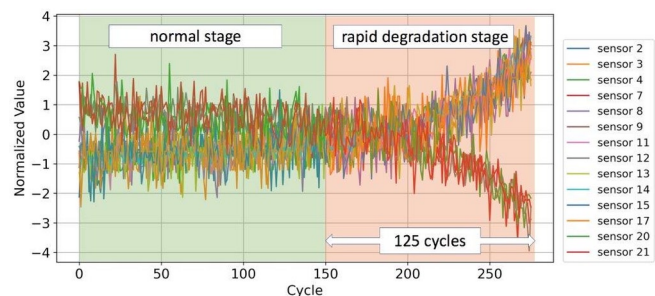


Fig. 8. Degradation failure threshold

Since the rapid degradation trend under normal stage is not obvious, the data intercepted before entering the rapid degradation stage is used as the training data, and the length of the data is set to 160 cycles.

To prevent the increase of network training difficulty caused by different sensor numerical scales, we need the z-score normalization for all the training data. The formula is shown as follows:

$$x'_i = \frac{x_i - u_i}{\sigma_i} \quad (23)$$

where u_i is the mean value and σ_i is the corresponding standard deviation.

4.2. Data Generated

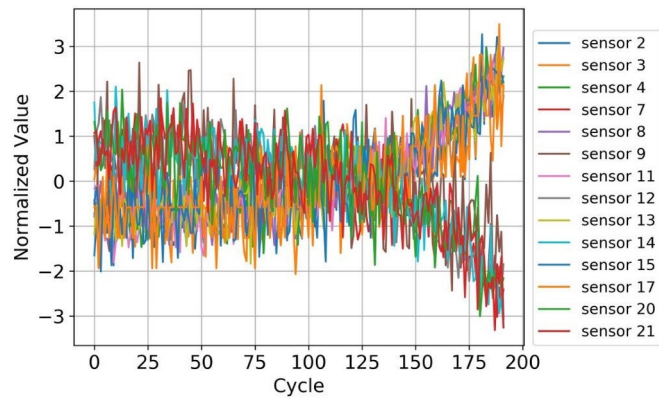
The preprocessed data are sent to the amplification network as input. Different from the regression task, the customized generator is a single-layer LSTM network that prevents the output of the network from becoming highly abstract and affecting the expression of the details of the original degradation data.

The number of parameters in LSTM has a positive correlation with the complexity of the model. In this experiment, the number of parameters in LSTM is set to 160. To keep the input and output dimensions of the network consistent, a dense operation is conducted at the output of the network. For the discriminator, features with degraded information need to be extracted extensively, so a stacked three-layer LSTM network is applied, and the number of parameters in LSTM is set to 100.

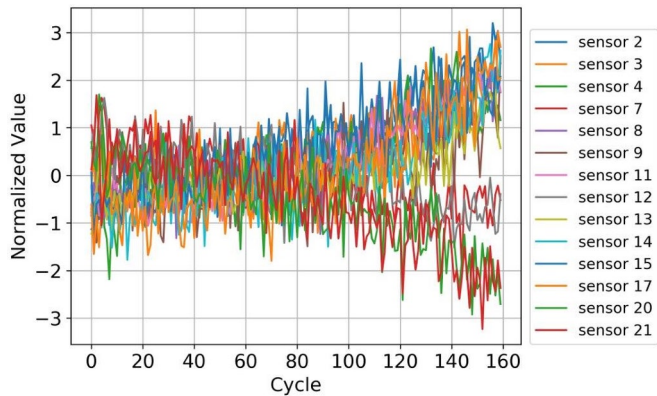
The trained generator from CycleGAN is used to amplify the training data. The data selected in this experiment are all from FD001. We divide 100 data into three groups of 7: 2: 1 as training set, validation set, and test set. In the training set, we use different numbers of data (10, 30, 50, and 70) to train the generator and explore the effect of various amounts of degradation data on the experimental results. The generators are constructed from different amounts of training data to generate FD001 Unit 1. The obtained data are shown in Figure 9. For further explanation, we number the data shown in Figure 9.

As shown in Figure 9, 1[#] indicates the original data, and 2[#], 3[#], 4[#], and 5[#] are the degradation data from the generator trained from the original data with different numbers of scales. In the case of the Generator built from less training data, such as Figure 9(b), the model can still learn the approximate distribution of samples. We compared the MMD differences between them, and the results are shown in Table III. Although these samples look similar from an intuitive point of view, they are not simply copied.

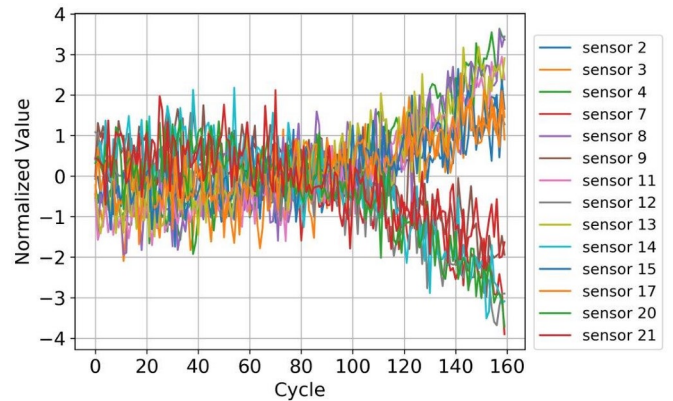
Furthermore, to find out the difference in the overall distribution of the generated degradation data, we compared the MMD between



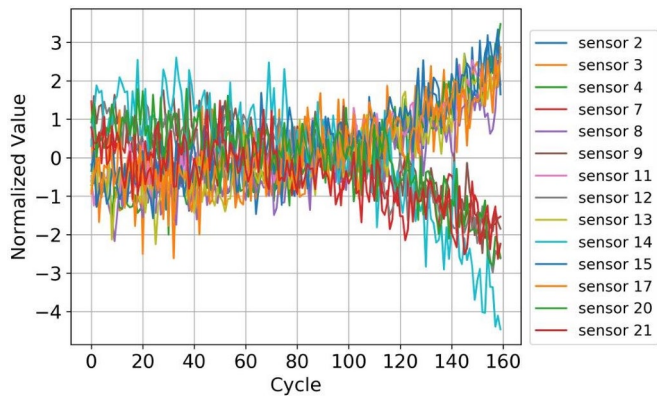
(a) 1[#]: FD001 Unit 1 for real



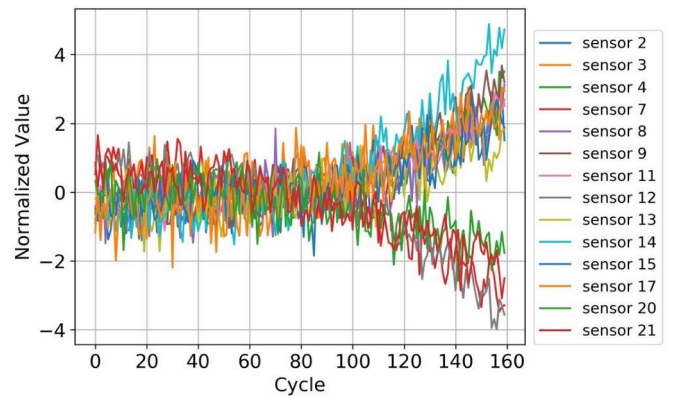
(b) 2[#]: FD001 Unit 1 generated by Generator trained with 10 samples



(c) 3[#]: FD001 Unit 1 generated by Generator trained with 30 samples



(d) 4[#]: FD001 Unit 1 generated by Generator trained with 50 samples



(e) 5[#]: FD001 Unit 1 generated by Generator trained with 70 samples

Fig. 9. FD001 Unit 1 generated from generator trained with different numbers of samples

a single generated sample in Figure 9 and all original training data shown in Figure 10.

Table III. MMD between real FD001 Unit 1 and generated FD001 Unit 1

| Data | 1 [#] | 2 [#] | 3 [#] | 4 [#] |
|------|----------------|----------------|----------------|----------------|
| MMD | 0.194 | 0.148 | 0.143 | 0.113 |

As the amount of data participating in training increases, the MMD between the generated and target samples is gradually reduced, which means that the generated samples and overall real samples are getting closer in distribution. Simultaneously, the trend of data degradation generated by the generator is more obvious, because the network summarizes the distribution of overall training data and provides the most common distribution. The generated degradation data are close to the real data in distribution. As the amount of data increases, the differ-

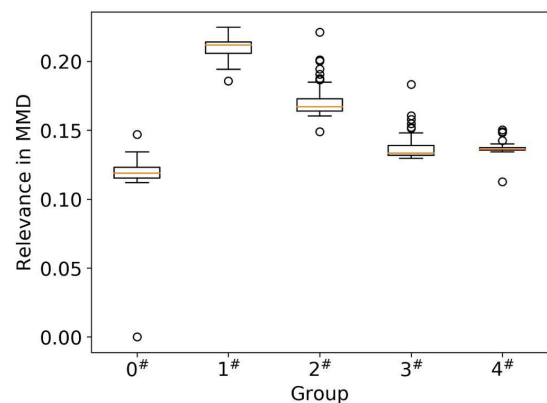


Fig. 10. MMD results of different groups. The figure shows a box plot of the MMD value between the FD001 Unit 1 generated by the generator constructed from different training samples and entire training samples

ence between the generated data and real data narrows, which is also in accordance with expectations.

C. RUL Prediction

To test whether the generated data can be used as training data to build a prediction network and explore the effect of our proposed model on training data of different sizes, we add the amplified data to the training data, and establish some prediction networks. To meet the requirements of controlled experiments, we built several sets of prediction models using real and generated data. The details are presented in the Table IV.

Table IV. Prediction network build with different data

| Group | General Method | Proposed Method |
|-------|-----------------|--------------------------------|
| A# | 10 real samples | 10 real + 10 generated samples |
| B# | 30 real samples | 30 real + 30 generated samples |
| C# | 50 real samples | 50 real + 50 generated samples |
| D# | 70 real samples | 70 real + 70 generated samples |

To accurately measure the prediction effect of the model, we present the evaluation method of RUL for the multi-sensor turbo aero engine as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N d_i^2} \quad (24)$$

where $d_i = RUL_i' - RUL_i$ indicates the prediction error of the i -th instance, RUL_i' and RUL_i respectively represent the predicted RUL from model and the actual RUL from dataset of the i -th instance. A score function is given as:

$$s = \sum_{i=1}^N s_i \quad (25)$$

$$s_i = \begin{cases} e^{\frac{d_i}{13}} - 1, & \text{for } |d_i| < 0 \\ e^{\frac{d_i}{10}} - 1, & \text{for } |d_i| \geq 0 \end{cases}$$

The score function from the dataset provider has a more practical significance. The penalty with smaller prediction deviations is small, but that for a larger prediction deviation is larger. The difference is shown in Figure 11.

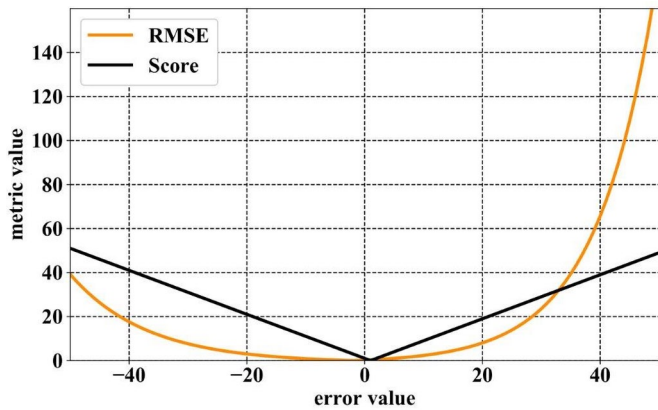


Fig. 11. Evaluation of RMSE and score function

The training data are grouped as A#, B#, C#, D#, and used to construct the RUL prediction model. To reduce the influence of the er-

ror on the experimental effect, each set of data builds a prediction model 10 times, and verifies the data on the test set. The average of the results is considered as the final results. The RMSE and scores are shown in Tables V and VI, and the results are plotted into the histogram in Figures 12 and 13.

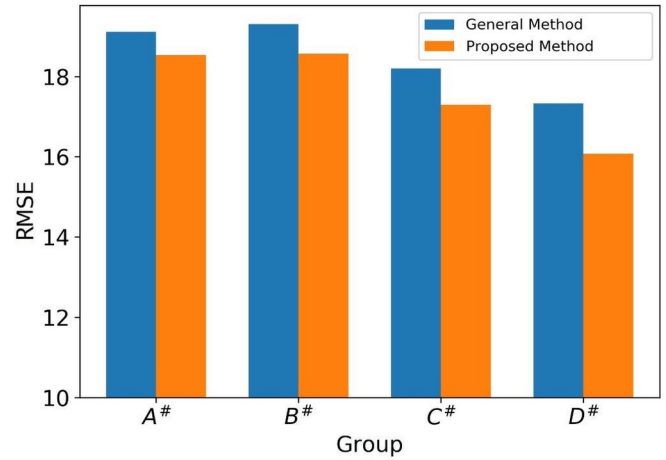


Fig. 12. RMSEs of the predicted RUL with the general method and proposed method

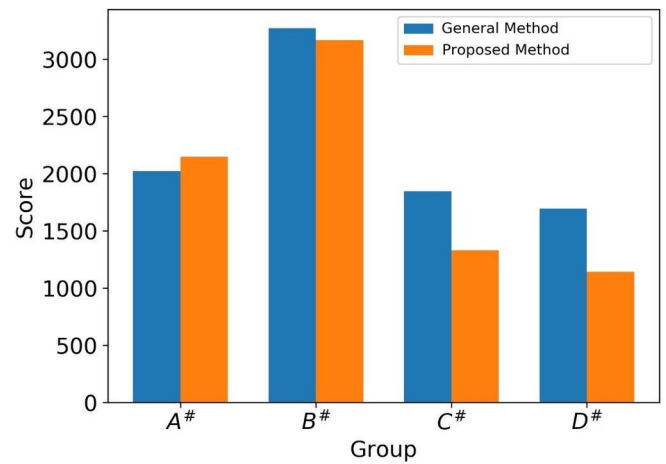


Fig. 13. Scores of the predicted RUL with the general method and proposed method

As shown in the Figure 12 and 13, the general method is reflecting on the blue histogram which is the result of a model built using the real data, and the proposed method is reflect on the orange histogram which is the result of a model built using not only the real data but also the generated data from the real data, what needs to be reminded is that both methods use the same predictive model, but the data used to build the model is different. When the MSE function is used to evaluate the test results, our proposed method achieves leading experimental results in all four groups of experiments. However, for the score function, the effect of groups A# and B# did not show obvious advantages, and the score of group B# is higher than that of group A#. In our analysis, the difference between the individual and test samples in the middle part of group B# is extremely large, and the score function is closer to the actual situation, resulting in the poor performance of the model built under the extremely small training data scale. When the real data increases, especially in the C# and D# groups, the proposed method performs better than the RUL prediction.

Intuitively, the RUL prediction in test set FD001 Unit 95 is shown in Figure 14. Sub-figures (a) to (d) indicate four results predicted by model constructed with real data. Sub-figures (e) to (h) indicate four

Table V. MSE of RUL results

| Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Average |
|----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|
| General Method / A# | 19.16 | 18.99 | 19.10 | 19.24 | 18.83 | 19.34 | 19.11 | 19.25 | 19.01 | 19.18 | 19.11 |
| Proposed Method / A# | 18.36 | 18.45 | 18.87 | 18.75 | 18.45 | 18.23 | 18.49 | 18.55 | 18.75 | 18.42 | 18.53 |
| General Method / B# | 19.92 | 19.64 | 19.03 | 19.87 | 18.88 | 19.08 | 19.20 | 18.95 | 19.21 | 19.22 | 19.30 |
| Proposed Method / B# | 18.67 | 19.20 | 18.54 | 17.05 | 18.74 | 19.03 | 18.26 | 19.09 | 19.08 | 18.06 | 18.57 |
| General Method / C# | 18.75 | 17.92 | 18.20 | 17.83 | 18.19 | 18.10 | 17.92 | 18.56 | 18.29 | 18.23 | 18.20 |
| Proposed Method / C# | 17.33 | 16.96 | 17.09 | 18.11 | 17.08 | 17.19 | 16.40 | 17.41 | 17.91 | 17.42 | 17.29 |
| General Method / D# | 16.77 | 17.34 | 18.47 | 17.10 | 17.32 | 18.02 | 17.09 | 16.26 | 18.09 | 16.86 | 17.33 |
| Proposed Method / D# | 17.05 | 16.33 | 15.66 | 15.11 | 16.44 | 15.58 | 16.24 | 16.07 | 15.57 | 16.60 | 16.07 |

Table VI. Score of RUL result

| Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Average |
|----------------------|------|------|------|------|------|------|------|------|------|------|---------|
| General Method / A# | 2148 | 1839 | 2134 | 2040 | 2014 | 2178 | 2077 | 1919 | 1865 | 2012 | 2023 |
| Proposed Method / A# | 1996 | 2963 | 1999 | 2152 | 1829 | 1605 | 1851 | 2879 | 2251 | 1965 | 2149 |
| General Method / B# | 3616 | 3393 | 2906 | 3556 | 2832 | 3729 | 3105 | 3339 | 2821 | 3408 | 3270 |
| Proposed Method / B# | 2632 | 3669 | 2793 | 2835 | 2498 | 2475 | 2510 | 3417 | 5610 | 3243 | 3168 |
| General Method / C# | 2324 | 1573 | 1729 | 1523 | 1947 | 2077 | 1566 | 1812 | 2058 | 1835 | 1844 |
| Proposed Method / C# | 1246 | 1295 | 1149 | 1677 | 1034 | 1294 | 924 | 1509 | 1554 | 1605 | 1329 |
| General Method / D# | 961 | 1940 | 2143 | 1467 | 2328 | 2205 | 1034 | 1631 | 1711 | 1519 | 1694 |
| Proposed Method / D# | 1218 | 1164 | 1141 | 922 | 1215 | 1192 | 1098 | 1281 | 1053 | 1138 | 1142 |

results predicted by model constructed with mixed data. Each result is predicted by a model built with different amounts of training data. Various amounts of training data can also build prediction networks, but the prediction effect constructed by the mixed data composed of real data and generated data is better. By comparing the results of MSE, we find that the curve convergence is better than the model built from real data. On the other hand, the prediction model with more training data has better model prediction effect, especially at the end of the life cycle, where the accuracy of the prediction is improved.

In the MSE evaluation, the proposed method can also improve the prediction accuracy. It is not obvious in the score evaluation of samples 10 and 30, but in samples 50 and 70, the proposed method has higher prediction scores.

4.3. Applicability analysis

The method proposed in this study is suitable for devices with multiple sensors and degradation data presented in time series. On the premise of having a small number of run-to-failure degradation data, our proposed method shows good performance, when a small amount of data is obtained, the remaining useful life of the equipment can also be effectively predicted. In the case of having sufficient degradation data, the sample space of the degradation data is sufficiently complete, and the prediction model established on this basis already has good performance, our proposed method has limited improvement under such circumstances. In view of the fact that obtain large amount of degradation data in actual industrial production is still not ideal, our proposed method still has very important significance.

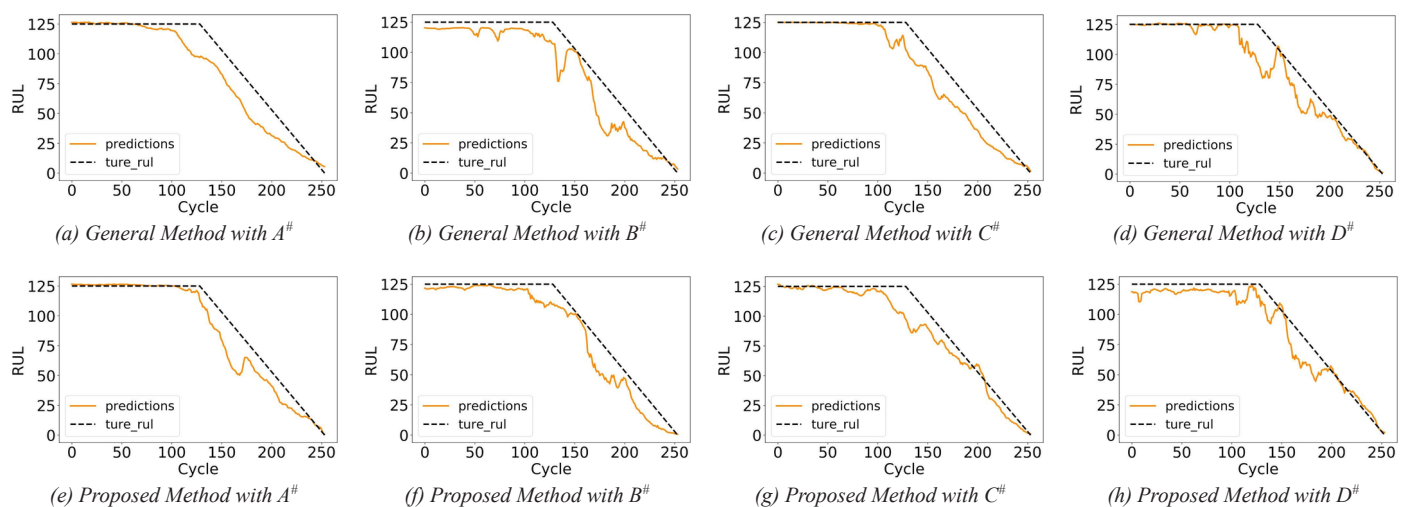


Fig. 14. RUL prediction results in the test set Unit 95

5. Conclusion

In this study, a framework for predicting the RUL with insufficient data was proposed, in which two main parts are involved. First, based on the characteristics of the sequence degradation data, an amplification network was designed using CycleGAN. Second, sliding time window strategy and deep BiLSTM network are jointly employed to construct the RUL prediction model based on the amplified degradation data. The following conclusions can be obtained: 1) Generating an adversarial network, as an unsupervised deep learning network, can indeed learn relevant information about data distribution. 2) The improved generated network based on LSTM can generate data with distribution similar to that of real data, and the RUL prediction network constructed using these amplified data has proved to be effective. 3) In the case where the RUL prediction accuracy is generally limited by the size of the training data, our proposed method provides a new reference for the development of RUL prediction.

References

1. Abdurraheem A, Abdullah Arshah R, Qin H. Evaluating the Effect of Dataset Size on Predictive Model Using Supervised Learning Technique. *International Journal of Software Engineering & Computer Sciences (IJSECS)* 2015; 1: 75–84, <https://doi.org/10.15282/ijsecs.1.2015.6.0006>.
2. Babu G S, Zhao P, Li X-L. Deep convolutional neural network based regression approach for estimation of remaining useful life. *International conference on database systems for advanced applications*, Springer: 2016: 214–228, https://doi.org/10.1007/978-3-319-32025-0_14.
3. Deutsch J, He D. Using deep learning-based approach to predict remaining useful life of rotating components. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2017; 48(1): 11–20, <https://doi.org/10.1109/TSMC.2017.2697842>.
4. Graves A, Mohamed A, Hinton G. Speech Recognition with Deep Recurrent Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2013. doi:10.1109/ICASSP.2013.6638947, <https://doi.org/10.1109/ICASSP.2013.6638947>.
5. Guo L, Li N, Jia F et al. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* 2017; 240: 98–109, <https://doi.org/10.1016/j.neucom.2017.02.045>.
6. KARABACAK E Yunus, GÜRSEL ÖZMEN N, GÜMÜŞEL L. Worm gear condition monitoring and fault detection from thermal images via deep learning method. *Eksplatacja i Niezawodność – Maintenance and Reliability* 2020; 22(3): 544–556, <http://dx.doi.org/10.17531/ein.2020.3.18>.
7. Khan S, Yairi T. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing* 2018; 107: 241–265, <https://doi.org/10.1016/j.ymssp.2017.11.024>.
8. Le Son K, Fouladirad M, Barros A et al. Remaining useful life estimation based on stochastic deterioration models: A comparative study. *Reliability Engineering & System Safety* 2013; 112: 165–175, <https://doi.org/10.1016/j.res.2012.11.022>.
9. Li D, Chen D, Jin B et al. MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. In Tetko IV, Kůrková V, Karpov P, Theis F (eds): *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, Cham, Springer International Publishing: 2019: 703–716, https://doi.org/10.1007/978-3-030-30490-4_56.
10. Li J, Li X, He D. A directed acyclic graph network combined with cnn and lstm for remaining useful life prediction. *IEEE Access* 2019; 7: 75464–75475, <https://doi.org/10.1109/ACCESS.2019.2919566>.
11. Li X, Ding Q, Sun J. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety* 2018; 172: 1–11, <https://doi.org/10.1016/j.res.2017.11.021>.
12. Li Y, Wang K. Modified convolutional neural network with global average pooling for intelligent fault diagnosis of industrial gearbox. *Eksplatacja i Niezawodność – Maintenance and Reliability* 2020; 22(1): 63–72, <http://dx.doi.org/10.17531/ein.2020.1.8>.
13. Lyu Y, Gao J, Chen C et al. Optimal Burn-in Strategy for High Reliable Products Using Convolutional Neural Network. *IEEE Access* 2019; 7: 178511–178521, <https://doi.org/10.1109/ACCESS.2019.2958570>.
14. Lyu Y, Gao J, Chen C et al. Joint model for residual life estimation based on Long-Short Term Memory network. *Neurocomputing* 2020; 410: 284–294, <https://doi.org/10.1016/j.neucom.2020.06.052>.
15. Mao W, He J, Zuo M J. Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. *IEEE Transactions on Instrumentation and Measurement* 2019. doi:10.1109/TIM.2019.2917735, <https://doi.org/10.1109/TIM.2019.2917735>.
16. Nieto P G, Garcia-Gonzalo E, Lasheras F S, de Cos Juez F J. Hybrid PSO–SVM-based method for forecasting of the remaining useful life for aircraft engines and evaluation of its reliability. *Reliability Engineering & System Safety* 2015; 138: 219–231, <https://doi.org/10.1016/j.res.2015.02.001>.
17. Peng K, Jiao R, Dong J, Pi Y. A deep belief network based health indicator construction and remaining useful life prediction using improved particle filter. *Neurocomputing* 2019; 361: 19–28, <https://doi.org/10.1016/j.neucom.2019.07.075>.
18. Ragab A, Ouali M-S, Yacout S, Osman H. Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation. *Journal of Intelligent Manufacturing* 2016; 27(5): 943–958, <https://doi.org/10.1007/s10845-014-0926-3>.
19. Ren L, Sun Y, Wang H, Zhang L. Prediction of bearing remaining useful life with deep convolution neural network. *IEEE Access* 2018; 6: 13041–13049, <https://doi.org/10.1109/ACCESS.2018.2804930>.
20. Sagheer A, Kotb M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* 2019; 323: 203–213, <https://doi.org/10.1016/j.neucom.2018.09.082>.
21. Si X-S, Wang W, Hu C-H et al. A Wiener-process-based degradation model with a recursive filter algorithm for remaining useful life estimation. *Mechanical Systems and Signal Processing* 2013; 35(1–2): 219–237, <https://doi.org/10.1016/j.ymssp.2012.08.016>.

Some possible topics for future research include the follows.

- (1) In many applications, the test set and training set may come from different test conditions, under which the equipment workloads, environmental condition and noise levels may vary. That may lead to different distribution of training set and test set. It would be interesting to improve the domain adaptability of our RUL prediction framework.
- (2) Due to the variability of raw materials quantity and manufacturing accuracy, it is common to see that the degradation characteristics of individuals may show unit-to-unit variability. How to improve prediction accuracy considering individual characteristics deserves further investigation.

Acknowledgements

This work was supported by the Major Special Projects of Zhongshan 200824103628344 and 2019A4018.

22. Sohani A, Sayyaadi H, Hoseinpoori S. Modeling and multi-objective optimization of an M-cycle cross-flow indirect evaporative cooler using the GMDH type neural network. *International Journal of Refrigeration* 2016; 69: 186–204, <https://doi.org/10.1016/j.ijrefrig.2016.05.011>.
23. Su C, Chen H, Wen Z. Prediction of remaining useful life for lithium-ion battery with multiple health indicators. *Eksploracja i Niezawodność – Maintenance and Reliability* 2021; 23(1): 176–183, <http://dx.doi.org/10.17531/ein.2021.1.18>.
24. Su C, Chen H, Wen Z. Prediction of remaining useful life for lithium-ion battery with multiple health indicators. *Eksploracja i Niezawodność – Maintenance and Reliability* 2021; 23(1): 176–183, <http://dx.doi.org/10.17531/ein.2021.1.18>.
25. Wang B, Lei Y, Li N, Li N. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability* 2018. doi:10.1109/TR.2018.2882682, <https://doi.org/10.1109/TR.2018.2882682>.
26. Wen B, Xiao X, Wang X et al. Data-driven remaining useful life prediction based on domain adaptation. *PeerJ Computer Science* 7:e690 2021. doi:<https://doi.org/10.7717/peerj-cs.690>, <https://doi.org/10.7717/peerj-cs.690>.
27. Xie Y, Zhang T. A transfer learning strategy for rotation machinery fault diagnosis based on cycle-consistent generative adversarial networks. 2018 Chinese Automation Congress (CAC), IEEE: 2018: 1309–1313, <https://doi.org/10.1109/CAC.2018.8623346>.
28. Yin S, Ding S X, Xie X, Luo H. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics* 2014; 61(11): 6418–6428, <https://doi.org/10.1109/TIE.2014.2301773>.
29. Yinka-Banjo C, Ugot O-A. A review of generative adversarial networks and its application in cybersecurity. *Artificial Intelligence Review* 2020; 53(3): 1721–1736, <https://doi.org/10.1007/s10462-019-09717-4>.
30. Yoon J, Drumright L N, van der Schaar M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics* 2020; 24(8): 2378–2388, <https://doi.org/10.1109/JBHI.2020.2980262>.
31. Zhai Q, Ye Z-S. RUL prediction of deteriorating products using an adaptive Wiener process model. *IEEE Transactions on Industrial Informatics* 2017; 13(6): 2911–2921, <https://doi.org/10.1109/TII.2017.2684821>.
32. Zhang X, Xiao P, Yang Y et al. Remaining Useful Life Estimation Using CNN-XGB with Extended Time Window. *IEEE Access* 2019; PP: 1–1, <https://doi.org/10.1109/ACCESS.2019.2942991>.
33. Zhang Y, Xiong R, He H, Pecht M G. Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. *IEEE Transactions on Vehicular Technology* 2018; 67(7): 5695–5705, <https://doi.org/10.1109/TVT.2018.2805189>.
34. Zheng G, Sun W, Zhang H et al. Tool wear condition monitoring in milling process based on data fusion enhanced long short-term memory network under different cutting conditions. *Eksploracja i Niezawodność – Maintenance and Reliability* 2021; 23(4): 612–618, <https://doi.org/10.17531/ein.2021.4.3>.
35. Zhu J, Chen N, Peng W. Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Transactions on Industrial Electronics* 2018; 66(4): 3208–3216, <https://doi.org/10.1109/TIE.2018.2844856>.
36. Zhu K, Liu T. Online tool wear monitoring via hidden semi-Markov model with dependent durations. *IEEE Transactions on Industrial Informatics* 2017; 14(1): 69–78, <https://doi.org/10.1109/TII.2017.2723943>.
37. Zio E, Di Maio F. A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system. *Reliability Engineering & System Safety* 2010; 95(1): 49–57, <https://doi.org/10.1016/j.ress.2009.08.001>.

A method for estimating the probability distribution of the lifetime for new technical equipment based on expert judgement

Indexed by:



Karol Andrzejczak^a, Lech Bukowski^b

^aPoznan University of Technology, Faculty of Control, Robotics and Electrical Engineering, Institute of Mathematics, ul. Piotrowo 3A, 60-965 Poznań, Poland

^bWSB University, ul. Zygmunta Cieplaka 1c, 41-300 Dabrowa Górnicza, Poland


Highlights

- A new method for estimating the probability distribution of the lifetime based on expert assessments is developed.
- The expert lifetime elicitation procedure is developed and applied to the Weibull lifetime.
- The quantile function is used to develop the expert method.
- The subjective Bayesian approach with models of classical probability theory is integrated.
- The objectification of the evaluation of experts to assign weights to their opinions is proposed.

Abstract

Managing the exploitation of technical equipment under conditions of uncertainty requires the use of probabilistic prediction models in the form of probability distributions of the lifetime of these objects. The parameters of these distributions are estimated with the use of statistical methods based on historical data about actual realizations of the lifetime of examined objects. However, when completely new solutions are introduced into service, such data are not available and the only possible method for the initial assessment of the expected lifetime of technical objects is expert methods. The aim of the study is to present a method for estimating the probability distribution of the lifetime for new technical facilities based on expert assessments of three parameters characterizing the expected lifetime of these objects. The method is based on a subjective Bayesian approach to the problem of randomness and integrated with models of classical probability theory. Due to its wide application in the field of maintenance of machinery and technical equipment, a Weibull model is proposed, and its possible practical applications are shown. A new method of expert elicitation of probabilities for any continuous random variable is developed. A general procedure for the application of this method is proposed and the individual steps of its implementation are discussed, as well as the mathematical models necessary for the estimation of the parameters of the probability distribution are presented. A practical example of the application of the developed method on specific numerical values is also presented.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

uncertainty, expert elicitation of lifetime, quantile function, Weibull distribution.

Acronyms

| | |
|-----|----------------------------------|
| CDF | Cumulative Distribution Function |
| ED | Expert Data |
| EEL | Expert Elicitation of Lifetime |
| ELV | Expanded Lower Value |
| ETD | Expanded Triangle Distribution |
| EUV | Expanded Upper Value |
| IRF | Invers Reliability Function |
| LF | Lifetime Family |
| PDF | Probability Density Function |
| REE | Reliability Engineer Expert |
| RF | Reliability Function |
| TD | Technical Device |

Notation

| | |
|-----------------|--------------------------------------|
| β | shape parameter |
| $\tilde{\beta}$ | shape parameter in the EEL procedure |
| $\bar{\beta}$ | aggregated shape parameter |
| η | scale parameter |
| $\tilde{\eta}$ | scale parameter in the EEL procedure |
| $\bar{\eta}$ | aggregated scale parameter |
| $\lambda(t)$ | failure rate function |
| $\Gamma(\cdot)$ | gamma function |
| $f(t)$ | PDF |
| $F(t)$ | CDF |
| k | number of experts |

(*) Corresponding author.

E-mail addresses: K. Andrzejczak - karol.andrzejczak@put.poznan.pl, L. Bukowski - lbukowski@wsb.edu.pl

| | |
|-------------------------------|---|
| $M_k(r_1; r_2)$ | matrix of the theoretical values of the location parameters |
| $\tilde{M}_k(r_1; r_2)$ | matrix of the expert location parameters |
| p | unreliability level |
| $\Pr(\cdot)$ | probability function |
| r | reliability level |
| $R(t)$ | RF |
| $R^{-1}(p)$ | IRF |
| t | exposure variable (e.g., time) |
| t_p | potential lifetime at the unreliability level p |
| $wbl(\beta)$ | one-parameter family of Weibull distributions |
| $wbl(\beta; \eta)$ | two-parameter family of Weibull distributions |
| $wbl((t_1, r_1), (t_2, r_2))$ | two-parameter Weibull distribution in the EEL parametrization |

1. Introduction

In today's increasingly competitive environment, designing and manufacturing reliable products is essential to the company's survival. An innovative reliability program for a manufacturing company can significantly improve the quality, performance and durability of a product, and ultimately the company's profitability and customer satisfaction. Reliability analysis of industrial equipment is one of the most dynamic branches of research and continues to be a challenge for many applications. For decades, statistical methods have been developed and used in reliability research, see, e.g., [1, 15, 24, 29, 31]. Software tools to support more and more complex reliability analyses are being developed, see, e.g., [16, 17, 18].

Nowadays, empirical statistical methods are supported by other methods. The Bayesian modelling framework is based on incorporation of different sources of quantitative and qualitative data in the model [4, 22, 37]. The article [8] concerns the estimation of low probabilities of failure in terms of structural reliability. Analytic models for predicting system lifetime are based on reliability block diagrams [22], fault trees [25], Markov chains, semi-Markov processes [14], stochastic Petri nets [10] or hierarchical models. Typically, such models capture uncertainty that is natural in the system being modelled. This includes random times to failure of components, random times for various recovery actions and randomness in the ability to detect a failure. The methodology of examining uncertainty in various aspects is presented in the articles [20, 33, 38]. Such uncertainty, known as aleatory uncertainty, is usually captured by beta, gamma, exponential, triangular, Weibull, lognormal, Bernoulli and other distributions. Computations and results obtained from such models thus account for the aleatory uncertainty in the system. Results of the model will depend upon the validity of the assumed distribution forms as well as the parameter values attached to these distributions. Assuming that the distribution forms are valid, parametric uncertainty is the subject of this paper.

The main challenge of fitting distribution to reliability data is finding the family of distribution and the values of the parameters that give the highest probability of producing the observed data. One of the most common probability density functions used in industry is the Weibull distribution [1]. The paper [2] gives an extensive review of some discrete and continuous versions of the modifications of the Weibull distribution.

Other concepts of uncertainty description are based on the notion of imperfect knowledge [9] and use methods beyond classical probability theory. Such concepts include methods of so-called generalized uncertainty [5], which also allow the use of expert knowledge based on data and information of an incomplete and sometimes ambiguous

nature. These methods provide opportunities for quantitative uncertainty assessment considering three main criteria, which can sometimes conflict with each other, namely:

- inclusion in the analysis and calculation of all verified data and information at the disposal of the expert,
- the abandonment of assumptions in the model which cannot be clearly and reliably justified,
- the orientation of the modelling process towards achieving the main objective, which is to develop an effective tool to support decision-making under uncertainty.

As the predominant type of uncertainty within this concept is epistemic uncertainty, the most used methods for its description are subjective probabilities (e.g., in the Bayesian approach) and the so-called imprecise probabilities (e.g., in the approach of fuzzy set theory).

In many industrial applications the basic criterion for the usability of a technical device is the quality of the product, which is a function of the technical condition of this device. However, in the case of other types of technical devices, such as e.g., infrastructural facilities, and especially of unique character, this methodology is not applicable. Our proposal concerns exactly such devices, for which it is not possible to obtain either direct – historical data, or indirect – data concerning the influence of the degradation of the examined device on the quality of the product.

The aim of this article is to present a method of estimating the lifetime probability distribution of new technical devices based on expert assessments of only a few parameters characterizing the expected lifetime of these objects. The method is based on a subjective Bayesian approach to the problem of randomness and integrated with models of classical probability theory. Due to its widespread use in maintenance of machinery and technical equipment, a Weibull model is proposed, and possible practical applications are shown for it.

This article is organized as follows. Section 2 presents a literature survey on the determination of subjective probability distributions based on expert opinion data. Special emphasis is placed on discussing methods that have been positively validated in so-called critical infrastructure (e.g., in risk analysis of dams). On this basis, and in particular the analysis of the strengths and weaknesses of these methods, a modified procedure for expert elicitation of probabilities for any continuous random variable, consisting of eight main steps is proposed in Section 3. A general procedure for applying this method is developed and the various steps in its implementation are discussed. Section 4 proposes a formal construction of the expert lifetime elicitation procedure and presents the mathematical models necessary to estimate the parameters of its distribution. Application of the Expert Elicitation of Lifetime (EEL) procedure to the Weibull lifetime distribution is the subject of Section 5. The next section presents a practical example of using the developed method on concrete numerical values. The article ends with a summary, conclusions and plans for further work within the ongoing research project.

2. Determination of subjective probability distribution based on expert judgement – literature review

The subjective probability should reflect a starting point of knowledge of an object of interest (so-called prior probability distribution), based on which a rational person would use Bayes' methodology, by means of new available information, to determine the modified probability distribution (so-called posterior probability distribution). Thus, this methodology is implemented in multiple steps; first the prior probability is elicited and then it is modified based on further available information.

The stimulus for the dynamic development of methods based on Bayesian inference has been the challenge of managing the risk of unitary systems with high levels of reliability and potentially high safety risks, such as reactors in the nuclear power industry. An example of an attempt to solve this problem can be found in the safety

study of nuclear reactors, concluded with a guide recommending the use of appropriate elicitation methods [36]. This type of methodology has also been used to assess environmental risks and their impact on the safety and health of whole populations as well as individual people [27].

As interest in this issue grew, more and more papers appeared in the field of psychology on human decision-making under uncertainty. The experiments generally consisted of asking questions to which the subjects did not know the answers, and then respondents were asked to quantify the degree of uncertainty in these responses. Mostly the psychologists who compiled the results of these studies assigned corresponding probabilities to the different degrees of uncertainty. As a result of this research, it was found that assessing the uncertainty of one's own knowledge tends to be subject to systematic errors, which were called biases. Galwey's publication [12] defines the most important of these biases, namely:

- accessibility - overestimating the chance of events that have happened recently and that we have easy access to in our memory,
- representativeness – assessing the chance of events based on irrelevant data, often incidentally linked to those events,
- anchoring – ignoring new data and information about events about which we have already formed an opinion, particularly in terms of the likelihood of their occurrence, and
- overconfidence – overestimating our knowledge and therefore underestimating the uncertainty of our assessment.

Until the early-1990s, assessments of these errors were descriptive based on widely accepted concepts presented in the work by Kahneman, Slovic, and Tversky [21]. In contrast, Morgan and Henrion's book [27] proposed a general procedure that could be used as a basis for developing a guide for performing rational elicitation. Summarizing the literature in this area, it can be stated that (based on [12]):

- the selection of experts should consider their technical, technological, managerial, and economic competence in the subject matter of the expert opinion, and ensure their independence from the owner of the object under assessment,
- elicitation should take place under the minimum constraints of both time and money, and should provide the experts with full access to all information on the object of the evaluation,
- the elicitation methodology should be carefully prepared before the experts start their work, and the experts should know and accept it,
- the entire elicitation process should be carefully and explicitly documented so that it can be reproduced in the future and its correctness and effectiveness critically analysed.

Current Best Practices by determination of subjective probability distribution based on expert judgement can be synthesized to the following procedure, which is based on several sources (e.g., [11, 12, 26, 27]):

- Using multiple experts, if possible, the more the better. It is particularly important to ensure that independent experts with in-depth knowledge and engineering experience participate in the elicitation.
- Asking experts not only about the expected or most likely value, but also about the smallest and largest possible values of the parameter being evaluated. It is recommended that the order of the questions should force the experts to first ask for the dispersion of the values of the parameter and only then for the expected value.
- Use of triangular decomposition for graphical description of elicitation results. In works [6] and [13] it is recommended to modify this distribution by assuming that it covers only 90% of the entire range of variability of the evaluated parameter. The remaining 10% can be distributed symmetrically between the lower and upper areas of variation of the parameter [6], or asymmetrically, with 2% around lower values and 8% around

upper values [13]. Figure 1 shows an example of the Expanded Triangle Distribution (ETD) concept (based on [7] and [12]).

- Some authors recommend that experts provide additional percentile values for the assessed parameter to verify the plausibility of the assessment and check its compliance with the assumed triangular distribution.
- Provide experts with the opportunity to access the results of the entire elicitation process and organise an additional session with all experts to critically analyse both the process procedure itself and its results.
- Documentation in full of all stages of the elicitation process, including a description of their progress, analysis of the results obtained and archiving of the whole so that each element of the process can be reproduced at any time in the future.

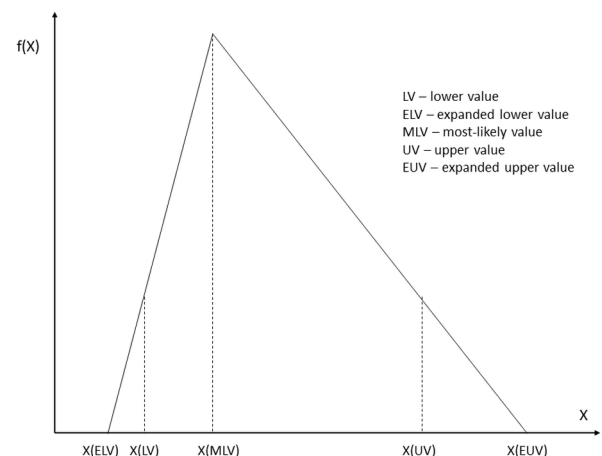


Fig. 1. The Expanded Triangle Distribution (ETD) concept – an example

Practical advice on the implementation of points b), c) and d) can be found e.g., in the publication on risk analysis of dams [32] in the form of suggested questions to experts:

- What is the lowest reasonably plausible number you can imagine the likelihood to be?
- What is the highest reasonably plausible number you can imagine the likelihood to be?
- Is it more likely to be somewhere in between these values?
- If so, what is the most likely value?
- The probability is not likely to be less than x ? (e.g., 10th percentile)
- The Probability is not likely to be more than y ? (e.g., 90th percentile)
- It cannot be less than v ? (e.g., 0th percentile) nor more than z ? (100th percentile)
- It is equally likely to be more or less than m ? (50th percentile)

The above-described methodology, based on the ETD concept, has been used successfully in several cases, e.g., in cost risk analysis [12]. However, in many cases, such as estimating the expected life of new technical facilities, it has proved unreliable. We see the main reasons for this situation in the following limitations of the ETD concept:

- The assumption that the range of a random variable X is restricted to a closed interval between ELV and EUV is contrary to maintenance experience on the durability of machinery and equipment.
- The values of 8 and 2% define the skewness of the probability distribution, but these are not universal values, and their adoption has not been sufficiently justified anywhere.
- In many practical situations it is crucial to determine probabilities for values of variable X outside the ELV to EUV range, which is impossible when using the ETD method.

In view of the above-mentioned limitations of the ETD method, the authors propose an alternative method devoid of these deficiencies. The assumptions of this method and the general procedure for its application is presented in Section 3.

3. Modified procedure for expert elicitation of a probability distribution for a continuous random variable

Based on the literature analysis conducted in Section 2 and our own experience, we propose a modified procedure for expert elicitation of a probability distribution for random variables, those of a continuous nature (e.g., expressed in units of time). The general procedure for the practical application of this method, consisting of eight steps, is shown in Figure 2.

Step one requires a clear, precise, and unambiguous formulation of the problem to be addressed by the experts. The experts should have all the relevant information for the evaluation, but not be burdened with unnecessary details that add little or nothing to the subject of the evaluation. The proper formulation of the task is the basis for the selection of appropriate experts who are authorities in the relevant field of knowledge.

The creation of as numerous and competent a group of experts is the objective of phase two. This is a difficult task, because usually these two criteria conflicts with each other – the more numerous the expert group, the greater the chance that it will also include less competent representatives. This step should also include selecting and adding to the expert team (or selecting from among them) an experienced facilitator, responsible for the harmonious work of the whole team – the group leader.

The next step is to develop an elicitation implementation plan, considering both organizational and scheduling aspects. All constraints (e.g., time, financial, etc.) should be considered, as well as possible disruptions that may occur during the elicitation process (e.g., threats and hazards). The plan should be as detailed as possible, but at the same time flexible (e.g., considering the possibility of one of the experts being indisposed). An important part of the plan is the preparation of appropriate forms for collecting data from experts, which should easily allow further computer processing of the information obtained.

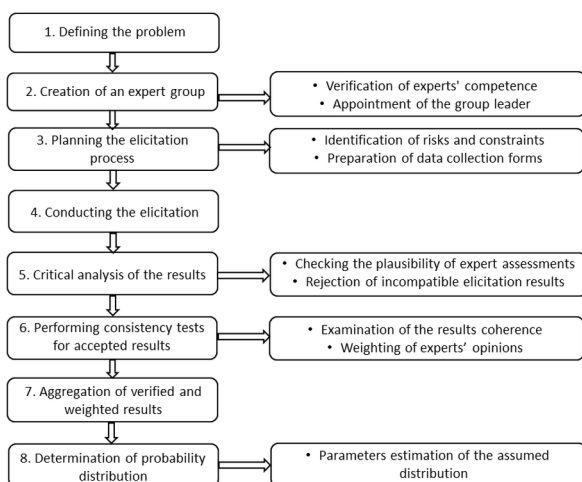


Fig. 2. General procedure for modified expert elicitation procedure of the lifetime distribution

Step four is a key part of the evaluation process, so it should proceed as quickly and smoothly as possible. To avoid possible mistakes of anchoring and suggesting the opinions of other team members each expert should perform the evaluation without contacting other experts.

The discussion on the assessment and arguing for or against certain opinions can take place in step five, after the work of step four

has been completely closed. In case of significant divergence between expert opinions, it is recommended to carry out an in-depth analysis, which should provide a conclusive answer to the question: are the results plausible? If the answer to this question is positive, you can proceed to step seven, which is to aggregate the results from all expert evaluators. If, on the other hand, the answer is negative, additional tasks must be taken to reach a compromise among the experts or to eliminate the opinions of those experts who could not convincingly justify their decisions.

In the first part of step six, additional verification of the consistency of the expert judgements should be carried out using a consistency test. The results of this test can be used as a basis for assessing the credibility of the individual experts and for assigning appropriate weights to their opinions in the second part of this step. This will allow the quality of individual elicitation to be considered in the process of aggregating the opinions of different experts.

The next step is to aggregate the verified elicitation results. The aggregation process uses the ratings of all the experts, considering the weights estimated in the previous step, in order to obtain unambiguous data allowing the estimation of the parameters of the assumed probability distribution.

The last step of the procedure is to create a parametric model of the lifetime probability distribution sought and to use it for practical purposes, e.g., determination of the expected lifetime of new technical equipment, for which the lack of operational data precludes the use of statistical methods.

The innovation of the proposed model is that the first 5 steps have been developed by modifying best practice in different areas of application of expert assessments used for critical infrastructures. The sixth and seventh steps, which aim to objectivize the assessments of individual experts, are fully innovative. We propose that verification of the consistency of the expert judgements should be carried out using a consistency test. The results of this test can be used as a basis for assessing the credibility of the individual experts and for assigning appropriate weights to their opinions in the second part of this step. This will allow the quality of individual elicitation to be considered in the process of aggregating the opinions of different experts. The aggregation process uses the ratings of all the experts, considering the weights estimated in the previous step, to obtain unambiguous data allowing the estimation of the parameters of the assumed probability distribution.

4. Formal construction of the expert lifetime elicitation procedure

We use the quantile method in the proposed procedure of the Expert Elicitation of Lifetime (EEL) of a Technical Device (TD). This method is often used in engineering research. For example, in the article [3], the quantile method was used to identify the costliest damage to parts of fleet vehicles. On the pages of Transport Topics [19] Evan Lockridge wrote "Engine makers are providing customers a gauge to help them determine how dependable and durable an engine is supposed to be, called a B-life rating." The construction of this lifetime measure is also based on a quantile function. The BX% rating in Weibull ++ is used to estimate the time when the probability of failure reaches a certain point (X%). Industry specialists consider this measure as a standard for measuring the life expectancy of technical products. For example, in predicting engine life, the most frequently heard ratings are B10% and B50% of life rating [19]. In this case B10% life is the expected engine durability expressed in kilometres of operation, before 10% of all operated engines of a specific type will require a major overhaul, renovation, or replacement. Thus, such information is very useful in giving customers a good idea of engine life expectations for a specific engine family. In practice BX% ratings are based on the durability data that engine manufacturers have on file and operating data [35]. So, a research problem appeared: *How to build an equivalent of this measure of lifetime for new technical*

devices for which operational data will appear only in the future? Our research is an attempt to solve this problem.

In our research, we do not have operational data or there is very little data, so we cannot use statistical methods to estimate parameters. Hence the need to develop an expert method for the assessment of unknown TD lifetime parameters. The primary role of the Reliability Engineering Expert (REE) is to identify hazards and manage the risks associated with the reliability of assets that may adversely affect the operations of a facility or company investing in new equipment. In such a case, we believe that the method of determining the lifetime of these equipment, developed in this article, may be useful. In the presented research, the BX% lifetime estimates are replaced with 100p% percentiles obtained from REE. Based on Expert Data (ED), the lifetime parameters of a predetermined family distributions are determined.

The likelihood of a system failure can be assessed under different circumstances using the REE group's opinion. It provides an applicable method for a facile computational prediction of future performances that aims to replace the usage of failure rates by a combination of instructed REE elicitation [28]. Due to the lack of historical data, expert judgment is used regarding the probability of the system failure in the planned operating conditions. Data on selected parameters of the lifetime are obtained using an appropriately designed questionnaire. In the designed survey, experts are asked to express their opinion on the potential lifetimes t_p at certain levels $p_1, \dots, p_l \in (0,1)$ of the unreliability in the assumed process of use and service for given TD. The originality of the developed lifetime parameter estimation procedure results from the application of this expert information for a specific lifetime model, instead of historical data. Such an approach to the issue of parameter evaluation has not yet been developed in the reliability theory.

Potential lifetime t_p at the unreliability level p is the quantile determined from the one of the equations $F(t_p) = p$ or $R(t_p) = 1 - p$. We assume that the potential lifetime is continuous, so t_p lifetime is derived from the quantile equation $t_p = R^{-1}(1 - p)$, where R^{-1} is the Invers Reliability Function (IRF). Potential lifetime t_p is the time during which the new TD will not fail with probability $r = 1 - p$. The potential lifetime t_p plays a fundamental role in developing the EEL procedure of TD. In the proposed EEL procedure, we use the fact that it is enough to know as many different potential lifetimes as there are parameters for the assumed Lifetime Family (LF) distributions. The characterization of the LF parameters of a given TD with the elaborated EEL procedure relies only on the potential lifetimes reported by a group of k independent REE experts.

Let TD be a new device (equipment) whose lifetime is to be estimated by a group of k REEs. Moreover, let $LF(\alpha_1, \dots, \alpha_s)$ denote the s parametric lifetime family of this device determined based on the knowledge of damage physics. The lack of historical data does not allow the use of statistical estimation of these parameters. In such a situation, we suggest using the EEL procedure to determine their value. As already indicated, the general idea of the expert elicitation is to use a potential lifetime. Experts from the REE group make individually elicitation the potential lifetimes t_{i1}, \dots, t_{is} for $i = 1, \dots, k$ and s different levels of reliability r_1, \dots, r_s or dually levels of unreliability $p_1 = 1 - r_1, \dots, p_r = 1 - r_s$. Moreover, they provide at least one location parameter as control values. Let l_{i1}, \dots, l_{iq} denote the control parameters of the i -th expert. The control parameters should be different from the selected potential lifetimes. Thus, we obtain ED as a two-block input matrix (1):

$$\tilde{M}_k(r_1; \dots; r_s) = \left[\begin{array}{ccc|ccc} \tilde{t}_{11} & \dots & \tilde{t}_{1s} & \tilde{l}_{11} & \dots & \tilde{l}_{1q} \\ \tilde{t}_{21} & \dots & \tilde{t}_{2s} & \tilde{l}_{21} & \dots & \tilde{l}_{2q} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \tilde{t}_{k1} & \dots & \tilde{t}_{ks} & \tilde{l}_{k1} & \dots & \tilde{l}_{kq} \end{array} \right] \quad (1)$$

Before we proceed to identifying lifetime distributions from the obtained ED, the group leader examines the plausibility of this data. At this stage, not plausible experts are rejected, and new experts are appointed in their place. The procedure is repeated until a fixed number of experts remain. The result of the work of the group leader is to establish a group of k experts and obtain an updated expert plausible data matrix $\tilde{M}_k(r_1; \dots; r_s)$. Only the data of the first block is needed to determine the LF parameters. The data contained in the second block we will use to determine weights for individual experts. To determine the LF parameters $\alpha_{i1}, \dots, \alpha_{is}$ for the i -th expert, a system of equations (2) is solved:

$$\begin{cases} R_{LF(\alpha_1, \dots, \alpha_s)}(\tilde{t}_{i1}) = r_1 \\ \dots & \text{for } i = 1, \dots, k \\ R_{LF(\alpha_1, \dots, \alpha_s)}(\tilde{t}_{is}) = r_s \end{cases} \quad (2)$$

where $R_{LF(\alpha_1, \dots, \alpha_s)}$ is RF of the $LF(\alpha_1, \dots, \alpha_s)$. If there exist Invers Reliability Function (IRF) $R_{LF(\alpha_1, \dots, \alpha_s)}^{-1}$, the parameters of potential expert lifetime can be determined by solving equivalent systems of equations (3):

$$\begin{cases} R_{LF(\alpha_1, \dots, \alpha_s)}^{-1}(r_1) = \tilde{t}_{i1} \\ \dots & \text{for } i = 1, \dots, k \\ R_{LF(\alpha_1, \dots, \alpha_s)}^{-1}(r_s) = \tilde{t}_{is} \end{cases} \quad (3)$$

Thus, for the i -th expert we obtain a random lifetime \tilde{T}_i , the probability distribution of which has the form (4):

$$\tilde{T}_i \sim LF(\tilde{\alpha}_{i1}, \dots, \tilde{\alpha}_{is}), i = 1, \dots, k \quad (4)$$

Based on the first block of the ED matrix, we obtained expert parameters $\tilde{\alpha}_{i1}, \dots, \tilde{\alpha}_{is}$ of the given LF distribution for all k experts. The obtained random lifetimes $\tilde{T}_1, \dots, \tilde{T}_k$ are necessary to perform consistency tests. In this step, we proceed to determine the theoretical values l_{i1}, \dots, l_{iq} of the control parameters for all k experts. In this way, we obtain the matrix (5) of the theoretical values of the control parameters for all k experts:

$$M_k = \begin{bmatrix} l_{11} & \dots & l_{1q} \\ l_{21} & \dots & l_{2q} \\ \dots & \dots & \dots \\ l_{k1} & \dots & l_{kq} \end{bmatrix} \quad (5)$$

The data consistency test is carried out for each expert separately. It consists in comparing the control parameters $\tilde{l}_{i1}, \dots, \tilde{l}_{iq}$ given by the i -th expert and recorded in the second block of the ED matrix, with their theoretical equivalents l_{i1}, \dots, l_{iq} determined from the obtained lifetimes $\tilde{T}_1, \dots, \tilde{T}_k$. If the control parameters given by a certain expert do not meet the conditions specified by the group leader, the data of that expert is omitted, and a new expert is appointed in his place.

If the ED matrix is plausible and consistent, then we proceed to the next step of the EEL procedure. In this step, based on the selected control parameter, the weights of the obtained lifetimes $\tilde{T}_1, \dots, \tilde{T}_k$ are determined. These weights are measures of the quality of the expert information contained in the ED matrix. The quality of the opinion of the i -th expert is assessed based on the relative measures of deviations dev_1 or dev_2 of the expert value $\tilde{\theta}$ of a given control parameter

from the theoretical value θ of this parameter. To determine the quality measures of expert opinions, we propose the formulas (6) and (7):

$$dev_1(\tilde{\theta}) \stackrel{\text{def}}{=} \frac{\tilde{\theta} - \theta}{\theta} \quad (6)$$

$$dev_2(\tilde{\theta}) \stackrel{\text{def}}{=} \frac{|\tilde{\theta} - \theta|}{\theta} \quad (7)$$

The obtained measures of relative deviations of expert values of control parameters from their theoretical values are used to determine the weights of the obtained lifetimes $\tilde{T}_1, \dots, \tilde{T}_k$. If $\tilde{\theta}_i \neq \theta$ for $i = 1, \dots, k$, then the weights are determined separately for the parameters as follows:

$$w_i(\tilde{\theta}) = \frac{1}{\sum_{j=1}^k \frac{1}{dev_2(\tilde{\theta}_j)}}, i = 1, \dots, k \quad (8)$$

If $\tilde{\theta}_i = \theta$ for a certain expert, then as the difference $|\tilde{\theta} - \theta|$ we take a small value, e.g., 0,000001. Then the obtained weights are used to determine the aggregated parameters $\tilde{\alpha}, \dots, \tilde{\alpha}_s$ of the TD lifetime \tilde{T} . Lifetime \tilde{T} parameterized in this way finalizes the presented EEL procedure, and its result is the weighted probability distribution (9):

$$\tilde{T} \sim \text{LF}(\tilde{\alpha}_1, \dots, \tilde{\alpha}_s) \quad (9)$$

The obtained lifetime \tilde{T} can be used to determine the functional and numerical both unconditional and conditional reliability characteristics of a TD.

However, it should be remembered that determining the LF parameters and its functional and numerical characteristics based on the EEL procedure is not always an easy task, as it may be necessary to know the specific properties of the families of lifetime distributions.

In the next section, we will do this for the family of Weibull lifetime distribution. Weibull lifetime can be applied to many situations. The main advantage of using this probability distribution is that it is flexible enough to accommodate different types of TD lifetimes and its well-known properties. Some of them that are useful for the EEL procedure are also presented in the next section.

5. Application of the EEL procedure to the Weibull lifetime distribution

Starting with a three-parameter Weibull lifetime distribution, the general Weibull model is given by the following Probability Density Function (PDF) [30]:

$$f_{wbl}(\beta; \eta; \gamma)(t) = \frac{\beta}{\eta} \left(\frac{t - \gamma}{\eta} \right)^{\beta-1} e^{-\left(\frac{t - \gamma}{\eta} \right)^\beta} \mathbb{I}_{[\gamma, \infty)}(t), \beta > 0, \eta > 0, -\infty < \gamma < \infty \quad (10)$$

where β is the shape parameter, η is the scale parameter, γ is the location parameter and $\mathbb{I}_{[\gamma, \infty)}$ is the indicator function (11):

$$\mathbb{I}_{[\gamma, \infty)}(t) := \begin{cases} 1 & \text{if } t \in (\gamma, \infty) \\ 0 & \text{if } t \notin (\gamma, \infty) \end{cases} \quad (11)$$

Since the main properties of the Weibull lifetime distribution is determined by the scale and shape parameters, we will focus further on the one- and two-parameter family of Weibull lifetime.

5.1. One-parameter Weibull lifetime distribution

This part of the publication presents the results of research on the properties of the Weibull distribution depending only on the shape parameter. These properties allow for a better eliciting information of the location characteristics and hence, the one-parameter Weibull lifetime plays a special role in our study. This special case occurs when the scale parameter is one and the location parameter is zero. In this case, one can only speak of a relative lifetime without entering unit names. PDF $f_{wbl}(\beta)$ for the one-parameter Weibull lifetime $wbl(\beta)$ reduces to (12):

$$f_{wbl}(\beta)(t) = \beta t^{\beta-1} e^{-t^\beta} \mathbb{I}_{[0, \infty)}(t), \beta > 0 \quad (12)$$

Now let's look at the effects of the beta shape parameter. The Fig. 3 shows the effect of different values of the shape parameter, β , on the shape of the PDF, independently of the other parameters. As you can see, the shape can take on a variety of forms based on the value of β .

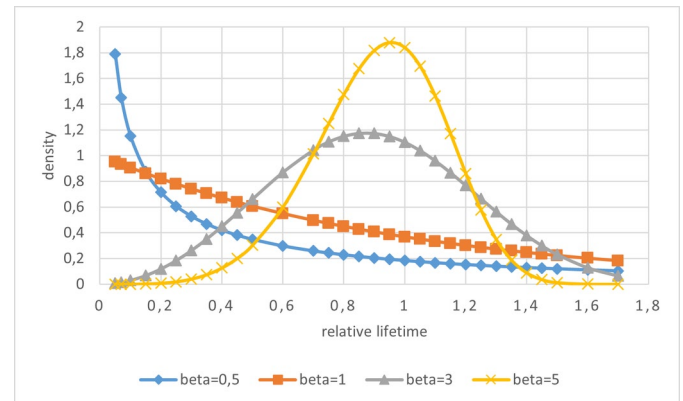


Fig. 3. One-parameter Weibull density curves for $\beta = 0,5$; 1; 3; and 5

As can be found in [34] for $\beta < 2,6$ the Weibull PDF is positively skewed, for $2,6 < \beta < 3,7$ coefficient of skewness approaches zero and consequently, it may approximate the normal PDF. For $\beta > 3,7$ it is negatively skewed. If $1 < \beta \leq 2$, then density function is concave downward and then upward, with inflection point given in (13):

$$t = \left(\frac{3(\beta - 1) + \sqrt{(5\beta - 1)(\beta - 1)}}{2\beta} \right)^{\frac{1}{\beta}} \quad (13)$$

If $\beta > 2$ density function is concave upward, then downward, then upward again, with inflection points at (14):

$$t = \left(\frac{3(\beta - 1) \pm \sqrt{(5\beta - 1)(\beta - 1)}}{2\beta} \right)^{\frac{1}{\beta}} \quad (14)$$

The Fig. 4 shows the effects of these varied values of β on the reliability plot. From the Fig. 4 it is clear, that all the reliability curves intersect at the point (1; 0,368). The following is the plot of the Weibull failure rate with the same values of β as above.

In Fig. 5 we can see that the failure rate can take various shapes informing about the type of aging of the TD. If $\beta > 2$, then the curve $\lambda(t)$ is convex and its slope increases with the increase of t . Conse-

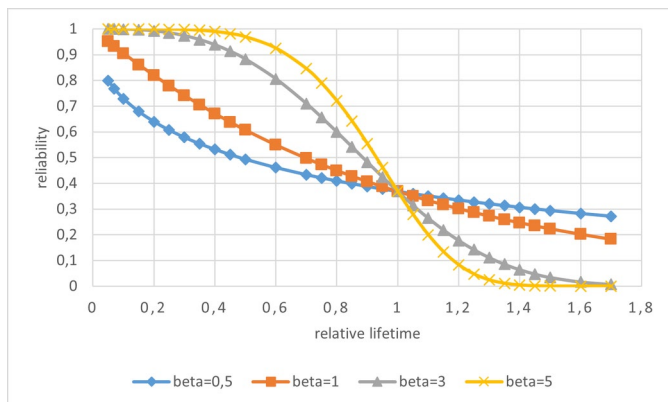


Fig. 4. One-parameter Weibull reliability curves for $\beta = 0,5$; 1; 3; and 5

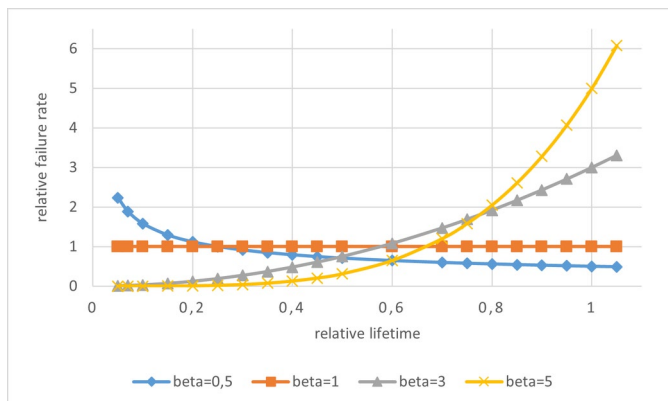


Fig. 5. One-parameter Weibull failure rate for $\beta = 0,5$; 1; 3; and 5

quently, the failure rate increases at an increasing rate as t increases, indicating wear out life. Depending on how skewness is measured we have different values of β giving a value of zero for the measure of skewness chosen [30]:

$\beta \approx 3,60235$ for skewness = zero,

$\beta \approx 3,43954$ for mean = median,

$\beta \approx 3,31247$ for mean = mode,

$\beta \approx 3,25889$ for mode = median.

Regarding the kurtosis, we have two values of β ($\beta \approx 2,25200$ and $\beta \approx 5,77278$) giving kurtosis = 3. The standardized normal and Weibull distributions have the same mean hazard rate = 0,90486 when $\beta \approx 3,43927$, which is nearly the value of shape parameter such that the mean is equal to the median. The effect of β can be translated into various modes of failures, as given in Table 1.

Table 1. Type of failures corresponding to β values

| β value | type of failure | meaning |
|-----------------|------------------|--|
| $\beta < 1$ | infant mortality | high probability of failing at early stages |
| $\beta = 1$ | random failures | failures are independent of time |
| $1 < \beta < 4$ | early wear out | can be due to generic failure modes, such as corrosion |
| $\beta \geq 4$ | rapid wear out | steep curve with fast wear out at some point |

As we can see, the shape parameter provides important information about the aging process of the TD for which we do not have statistical data yet. Determination of this parameter based on ED plays a key role in predictive research.

5.2. Two-parameter Weibull lifetime distribution

We now assume that the expert elicitation of the potential lifetimes refers to TD, whose lifetime T belongs to the two-parameter family

of Weibull distributions $wbl(\beta; \eta)$, where β is the shape parameter and η is the scale parameter. The Weibull lifetime with its two parameters permits the modelling of different regions of the bathtub curve in the lifecycle of a great number of components [37]. PDF $f_{wbl}(\beta; \eta)$ of the two-parameter Weibull's lifetime takes the form (15):

$$f_{wbl}(\beta; \eta)(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta} \right)^{\beta-1} e^{-\left(\frac{t}{\eta} \right)^{\beta}} \mathbb{I}_{(0, \infty)}(t), \quad \beta > 0, \eta > 0 \quad (15)$$

Scale parameter η is life characteristic because it is the time T such that $\Pr(T \leq \eta) = 0,632$. For two-parameter family $wbl(\beta; \eta)$, if $\beta = 1$, then failure rate is constant $\lambda_{wbl(1; \eta)}(t) = \frac{1}{\eta}$ and LF $wbl(1; \eta)$ is the exponential LF. For $\beta = 2$ the family $wbl(2; \eta)$ is the Rayleigh LF with a linearly increasing failure rate. For $0 < \beta < 1$ Weibull lifetime are characterized by decreasing failure rate. Thus, depending on the shape parameter, the Weibull distribution belongs to one of the classes: IFR, DFR or CFR, denoting, respectively, classes of increasing, decreasing or constant failure rate. For more details on this distribution and application, see the work of [30].

Potential lifetime t_p of the Weibull lifetime in engineering terminology defined as [23] takes the form (16):

$$t_p = \eta \cdot \left(\ln \left(\frac{1}{1-p} \right) \right)^{\frac{1}{\beta}}, \quad 0 < p < 1 \quad (16)$$

The experts' task is to assess potential lifetimes t_p for two given probability levels $p_1, p_2 \in (0,1)$. The data comes from the k REE group with comparable knowledge and sufficient experience in the management, maintenance, and design departments. As opinions differ, aggregation is performed to produce a single Weibull lifetime model. For this purpose, a weighting factor is calculated for each expert so that a weighted average of the opinions can be calculated. In summary, the steps to be taken to create an effective aggregate potential lifetime \tilde{T} of a TD using the EEL procedure for family $wbl(\beta; \eta)$ are as follows:

- Appointment of a group of k experts and a group leader to assess the durability of a new TD designed to operate under established operating conditions.
- Obtaining a plausible and consistent ED matrix of input data composed of potential lifetimes $\tilde{t}_{i,1}, \tilde{t}_{i,2}$ for two reliability levels $r_1 = 1 - p_1$, $r_2 = 1 - p_2$ and additional location parameters $\tilde{l}_{i,1}, \dots, \tilde{l}_{i,q}$ for control purposes.
- Determination of Weibull's lifetime $\tilde{T}_i \sim wbl(\tilde{\beta}_i; \tilde{\eta}_i)$ of the i -th expert, for $i = 1, \dots, k$.
- Calculation of the theoretical values of the control parameters $\tilde{l}_{i,1}, \dots, \tilde{l}_{i,q}$ for the obtained expert lifetime $\tilde{T}_i, i = 1, \dots, k$. The control parameters can be a mode, median, expected value, or other numeric localization measures.
- Selection of a control parameter as a weighting criterion and calculation of weights for individual expert opinions.
- Determination of the weighted Weibull potential lifetime $\tilde{T} \sim wbl(\tilde{\beta}; \tilde{\eta})$ for the selected criterion and two different reliability levels r_1, r_2 .
- Finally, it remains to use the obtained lifetime \tilde{T} to calculate the unconditional or conditional probabilities of survival of the TD and its functional and numerical characteristics useful in reliability tests.

Using the presented EEL procedure for determining the aggregated lifetime, we move to the formal calculation side. Let t_{i1} and t_{i2} for

$i = 1, \dots, k$ be given the potential lifetimes for two different reliability levels r_1 and r_2 for the TD starting the mission at age zero be given. To determine the parameters $\tilde{\beta}_i$ and $\tilde{\eta}_i$ for the ED of the i -th expert, system of equations (17) should be solved:

$$\begin{cases} \tilde{\eta}_i (-\ln(r_1))^{\frac{1}{\tilde{\beta}_i}} = t_{i1} \\ \tilde{\eta}_i (-\ln(r_2))^{\frac{1}{\tilde{\beta}_i}} = t_{i2} \end{cases} \quad (17)$$

The aim is to determine the parameters $\tilde{\beta}_i$ and $\tilde{\eta}_i$ of the Weibull's lifetime \tilde{T}_i as a function of the pairs (t_{i1}, r_1) and (t_{i2}, r_2) given by i -th expert. Solving the system (17) due to the scale parameter we get (18):

$$\begin{cases} \tilde{\eta}_i = \frac{t_{i1}}{(-\ln(r_1))^{\frac{1}{\tilde{\beta}_i}}} \\ \tilde{\eta}_i = \frac{t_{i2}}{(-\ln(r_2))^{\frac{1}{\tilde{\beta}_i}}} \end{cases} \quad (18)$$

After comparing the right sides of the (18), we get an equation with one unknown parameter $\tilde{\beta}_i$, which can be expressed as a function of the variables (t_{i1}, p_1) and (t_{i2}, p_2) . Thus, the solution (19) for the parameter $\tilde{\beta}_i$ is obtained as a function of (t_{i1}, r_1) , (t_{i2}, r_2) :

$$\tilde{\beta}_i = \log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right) \quad (19)$$

By inserting the determined shape parameter $\tilde{\beta}_i$ into the first equation (18) we get scale parameter $\tilde{\eta}_i$:

$$\tilde{\eta}_i = \frac{t_{i1}}{(-\ln(r_1))^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} \quad (20)$$

The lifetime distribution $wbl(\tilde{\beta}_i; \tilde{\eta}_i)$ determined in this way is an expert distribution of the two-parameter Weibull lifetime \tilde{T}_i . The random lifetime \tilde{T}_i obtained by the EEL procedure is denoted by $\tilde{T}_i \sim wbl((t_{i1}, r_1), (t_{i2}, r_2))$. For the obtained \tilde{T}_i , its functional and numerical characteristics can be determined. In such parameterization, the RF takes the form (21):

$$R_{wbl((t_{i1}, r_1), (t_{i2}, r_2))}(t) = e^{-\left(\frac{t (-\ln(r_1))^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}}{t_{i1}} \right)^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} \quad (21)$$

Weibull's potential lifetime $t_{i,p}$, for $i=1, \dots, k$, $r \in (0,1)$ and $p=1-r$ takes the form (22):

$$t_{i,p} = \frac{t_{i1}}{(-\ln(r_1))^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} (-\ln(r))^{\frac{1}{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} \quad (22)$$

Thus, for the Weibull's potential lifetime, having the ED in the form $(t_{i1}, r_1), (t_{i2}, r_2)$, it is possible to determine the scale parameter $\tilde{\eta}_i$, and the shape parameter $\tilde{\beta}_i$, and then calculate the lifetime location parameters for the i -th expert, such as: expected value (ev), mode (mo) and quartiles, in particular the median (me) and measure of deviation or skewness. Of course, having an ED, we can directly use it to calculate these measures. Apart from the expert's index, the calculation formulas for them take the form (23), (24) or (25), respectively:

$$ev(wbl((t_{i1}, r_1), (t_{i2}, r_2))) = \frac{t_{i1}}{(-\ln(r_1))^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} \Gamma \left(1 + \frac{1}{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)} \right) \quad (23)$$

$$mo(wbl((t_{i1}, r_1), (t_{i2}, r_2))) = \frac{t_{i1}}{(-\ln(r_1))^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} \left(\frac{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right) - 1}{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)} \right)^{-1} \quad (24)$$

$$me(wbl((t_{i1}, r_1), (t_{i2}, r_2))) = \frac{t_{i1}}{(-\ln(r_1))^{\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right)}} (\ln 2)^{\left(\log_{\frac{t_{i1}}{t_{i2}}} \left(\frac{\ln(r_1)}{\ln(r_2)} \right) \right)^{-1}} \quad (25)$$

These are the potential localization characteristics that are used in this article to construct quality measures of the EEL by comparing REE characteristics with their theoretical counterparts. The resulting Weibull's lifetime is used to determine the potential lifetime for a given reliability level r . Of course, complementary probability $p=1-r$ is the unreliability with which we want to determine the potential lifetime. The potential life t_p is calculated from the equation $R_{wbl(\beta; \eta; \gamma)}(t_p) = 1-p$. The lifetime t_p at the percentile level $100(1-p)\%$ denotes that the TD will be operational during this time at the reliability $r=1-p$. For example, $t_{0,1}$ is the lifetime at which given TD will be operational with the probability 0,9. Fig. 6 shows the relationship between potential lifetime and shape parameter β for various values of risk level p ($p=0,02; 0,04; 0,06; 0,08; 0,10$) and scale parameter $\eta=3500$. For $p=0,02$ potential life $t_{0,02}$ is the lifetime counted in adopted units of time, for which the TD will have a failure probability of 0,02.

Larger the value of β , longer the potential lifetime for the same value of η . In the presented probabilistic concept of determining Weibull distribution parameters, the potential lifetime t_p of the TD for a given probability level p is assessed by experts, because of their specialist knowledge.

6. Exemplification of the presented EEL procedure

Assuming that, the lifetime of the tested TD belongs to the family $wbl(\beta; \eta)$, k REE assess potential lifetimes t_1 and t_2 for two different reliability levels $r_1, r_2 \in (0,1)$ and basic location parameters: modal value mo , median me and expected value ev . Thus, the ED received from REE takes the form of the mapping (26):

$$ED: (0,1)^2 \ni (r_1, r_2) \rightarrow (\tilde{t}_1, \tilde{t}_2 | \widetilde{mo}, \widetilde{me}, \widetilde{ev}) \in \mathbb{R}_+^5 \quad (26)$$

where \mathbb{R}_+^5 denotes the Cartesian product of positive real numbers.

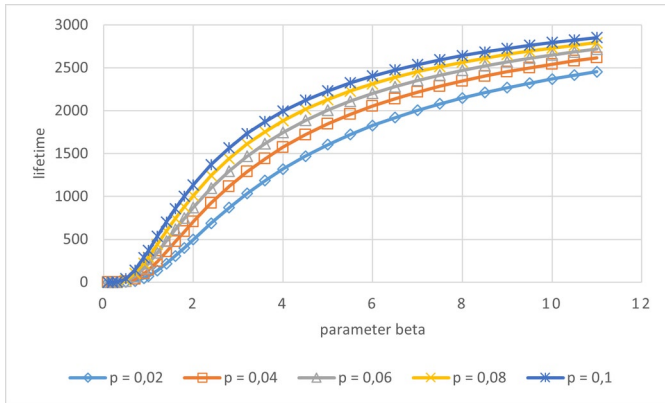


Fig. 6. Potential lifetime t_p versus β for $\eta = 3500$ and $p = 0,02$ (bright blue), $p = 0,04$ (light brown), $p = 0,06$ (gray), $p = 0,08$ (yellow), $p = 0,1$ (dark blue)

If k experts evaluate the location parameters of the potential lifetime of the same TD, based on the same two reliability levels r_1, r_2 , then we obtain the set of ED in the form of five-dimensional vectors arranged in the matrix \tilde{M}_k (27):

$$\tilde{M}_k(r_1; r_2) = \begin{bmatrix} \tilde{t}_{11} & \tilde{t}_{12} & \tilde{m}o_1 & \tilde{m}e_1 & \tilde{e}v_1 \\ \tilde{t}_{21} & \tilde{t}_{22} & \tilde{m}o_2 & \tilde{m}e_2 & \tilde{e}v_2 \\ \dots & \dots & \dots & \dots & \dots \\ \tilde{t}_{k1} & \tilde{t}_{k2} & \tilde{m}o_k & \tilde{m}e_k & \tilde{e}v_k \end{bmatrix} \quad (27)$$

The theoretical values of the location parameters are determined based on two potential lifetimes \tilde{t}_{i1} and \tilde{t}_{i2} , given by REE for $i = 1, \dots, k$. In this way we obtain a matrix M_k of the theoretical values of the lifetime location parameters $m o_i, m e_i, e v_i$:

$$M_k(r_1; r_2) = \begin{bmatrix} m o_1 & m e_1 & e v_1 \\ m o_2 & m e_2 & e v_2 \\ \dots & \dots & \dots \\ m o_k & m e_k & e v_k \end{bmatrix} \quad (28)$$

Let's illustrate these matrices with a practical example for given reliability level $r_1 = 0,9$ and $r_2 = 0,1$. The opinions of the group of $k = 4$ REE on potential lifetime parameters for a certain TD used continuously, presented in the form of a matrix $M_4(0,9;0,1)$, are as (29), where the unit of time is the one day of using TD:

$$\tilde{M}_4(0,9;0,1) = \begin{bmatrix} 3500 & 4500 & 4000 & 4000 & 4000 \\ 3200 & 4800 & 4000 & 4000 & 4000 \\ 3000 & 4500 & 3500 & 3500 & 3500 \\ 2800 & 4000 & 3500 & 3500 & 3500 \end{bmatrix} \quad (29)$$

To assess the quality of the ED, we calculate the theoretical values of the control parameters for all four experts. This is how we get the matrix (30):

$$M_4(0,9;0,1) = \begin{bmatrix} 4175 & 4081 & 4032 \\ 4223 & 4099 & 4041 \\ 3959 & 3843 & 3788 \\ 3581 & 3843 & 3433 \end{bmatrix} \quad (30)$$

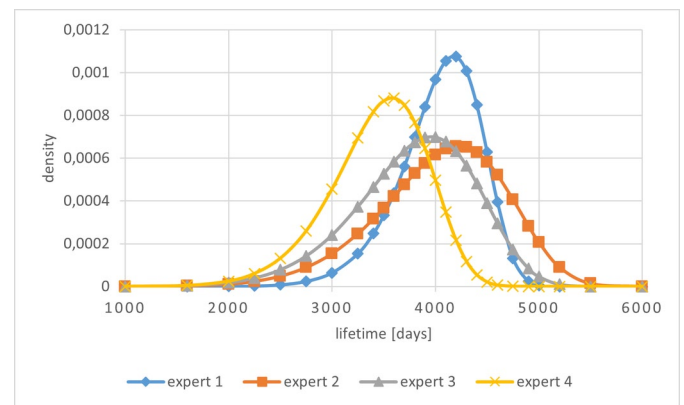
The Table 2 summarizes the parameters and potential lifetime t_p at the unreliability level $p = 0,01$ of the Weibull distribution determined for the given ED.

In all four cases, the beta parameter is greater than 4, which proves that all experts treated the tested TD in the same way as a high-quality object whose rapid wear occurs only after a longer period of use. For a graphical comparison, graphs of PDF curves (Fig. 7), reliability function (Fig. 8) and failure rate function (Fig. 9) were prepared for the obtained four expert Weibull lifetimes $\tilde{T}_1, \tilde{T}_2, \tilde{T}_3, \tilde{T}_4$.

Fig. 7. Two-parameter Weibull PDF curves for the first (blue), second (orange), third (gray) and fourth (yellow) expert

Figure 7 shows that the expert lifetimes are generally similar and Table 2. Parameters of the Weibull distribution determined for given ED

| Expert number | Shape parameter β | Scale parameter η | $t_{0,01}$ [days] |
|---------------|-------------------------|------------------------|-------------------|
| 1 | 12,273071 | 4204,35587 | 2890 |
| 2 | 7,607066 | 4301,55567 | 2350 |
| 3 | 7,607066 | 4032,708436 | 2203 |
| 4 | 8,647649 | 3632,23519 | 2134 |



are almost completely concentrated in the range of 1600 to 5600 days. As for Weibull distributions, they are characterized by high symmetry. This is due to the high values of the shape parameter. The mode values of the obtained lifetimes differ the most for the second and fourth experts, the difference being around 600 days. The lifetime by the first expert has the lowest dispersion, and the second and third experts have the greatest dispersion.

The presented graphs of the reliability function illustrate the differences of expert predictions. As can be seen from Fig. 8, the first and fourth expert are characterized by the maximum difference in the reliability value. This difference is achieved at 4000 days and

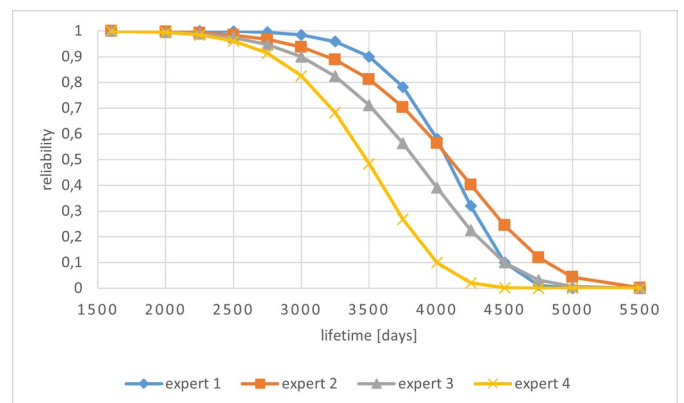
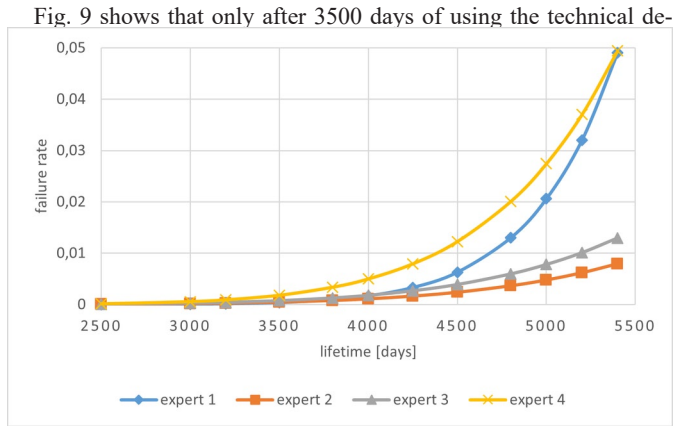


Fig. 8. Two-parameter Weibull RF for the first (blue), second (orange), third (gray) and fourth (yellow) expert

is approximately 0,5, but for 3000 and fewer days, this difference is already below 0,2.

The last presented function for individual experts is the failure rate (Fig. 9). This function at any time characterizes the relative deterioration of the reliability of the TD per day. In engineering practice, historical data on the device or system under consideration is traditionally used to determine this function. Here we showed how to derive this function based on the ED. In all cases, the $\lambda(t)$ curves are convex, and their slopes increase with the increase of t . Consequently, the failure rates increase with the increase of t , which additionally indicates the wear of the TD.

Fig. 9. Two-parameter Weibull failure rate for the first (blue), second (orange), third (gray) and fourth (yellow) expert



vice, the failure rates for all experts are greater than 0,001, and then their growth significantly accelerates. The greatest increase results from the data obtained from the first and fourth experts.

In the EEL procedure, we propose that the quality of the i -th expert eliciting information should be assessed based on relative deviation measure dev_1 of the expert value of control parameters, i.e., the mode, the median or the expected value from their theoretical values. Calculation results are summarized in the Table 3.

Table 3 shows that for the experts' elicitation based on the modal value, all four opinions were slightly underestimated and the opinion of the fourth expert was rated the highest. The fourth expert is also rated the highest in the median criterion, and this time this expert was

Table 3. Expert deviation for the first measure of deviation

| Expert number | $dev_1(\widetilde{mo})$ | $dev_1(\widetilde{me})$ | $dev_1(\widetilde{ev})$ |
|---------------|-------------------------|-------------------------|-------------------------|
| 1 | -0,04199 | -0,01977 | -0,00802 |
| 2 | -0,05271 | -0,02420 | -0,01013 |
| 3 | -0,11587 | -0,08926 | -0,07612 |
| 4 | -0,02261 | 0,005312 | 0,019426 |

the only one to provide a minimally overestimated value. In the case of the expected value criterion, except the fourth expert, the other experts again slightly lowered the expected value, and the first expert assessed this value most accurately.

The measure dev_2 of relative deviations of expert values of control parameters from their theoretical values are used to determine the weights of individual experts. The results of the weight calculations for all experts are presented in the Table 4.

The calculated weights of expert lifetime assessments confirm the expert opinion quality ranking. Taking a specific location parameter as a criterion, the obtained weights are used to calculate the aggregated shape $\tilde{\beta}$ and scale $\tilde{\eta}$ parameters. The calculation results are presented in the Table 5.

In this way, using the EEL procedure, we obtained the following aggregated lifetime distribution for the individual criteria:

Table 4. Weights of ED for individual location parameters

| Expert number | $w_i(\widetilde{mo})$ | $w_i(\widetilde{me})$ | $w_i(\widetilde{ev})$ |
|---------------|-----------------------|-----------------------|-----------------------|
| 1 | 0,25 | 0,17 | 0,43 |
| 2 | 0,20 | 0,14 | 0,34 |
| 3 | 0,09 | 0,04 | 0,05 |
| 4 | 0,46 | 0,65 | 0,18 |

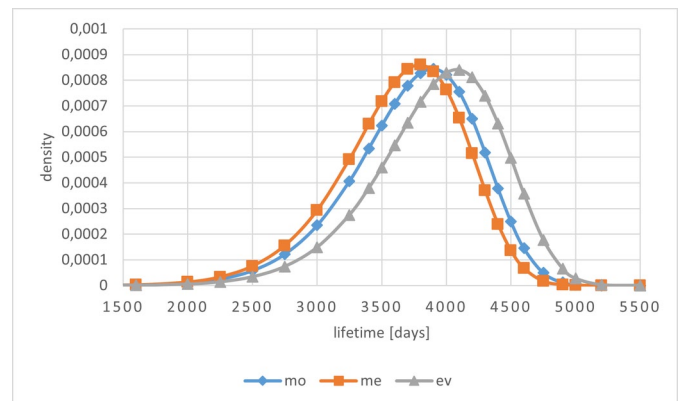
Table 5. List of the aggregated lifetime parameters for three criteria

| Characteristics | Criterion mo | Criterion me | Criterion ev |
|---------------------------------|----------------|----------------|----------------|
| Shape parameter $\tilde{\beta}$ | 8,998 | 8,923 | 9,386 |
| Scale parameter $\tilde{\eta}$ | 3945 | 3843 | 4129 |

$$\begin{cases} \tilde{T}_{mo} \sim wbl(8,998;3945) \\ \tilde{T}_{me} \sim wbl(8,923;3843) \\ \tilde{T}_{ev} \sim wbl(9,386;4129) \end{cases} \quad (31)$$

Density curves for the obtained aggregate distributions are presented in Fig. 10.

Fig. 10. Two-parameter aggregated Weibull density curves for the mo (blue), me (orange) and ev (gray) criterion



As can be seen from Fig. 10, the differences between the obtained distributions are relatively small. If we take the centrally located density curve as the criterion for selecting the aggregate lifetime, then in this case the lifetime mode should be selected.

Then, for the obtained aggregate lifetimes $\tilde{T}_{mo}, \tilde{T}_{me}, \tilde{T}_{ev}$ the mode, the median and the expected value were calculated from the formulas (32), (33) and (34), and for the three criteria under consideration.

$$mo(T) = \eta \left(1 - \frac{1}{\beta} \right)^{\frac{1}{\beta}}, \beta > 1 \quad (32)$$

$$me(T) = \eta (\ln 2)^{\frac{1}{\beta}} \quad (33)$$

$$ev(T) = \eta \Gamma\left(1 + \frac{1}{\beta}\right) \quad (34)$$

The results of the calculations are presented in the Table 6.

As would be expected for the mode criterion, we obtained the intermediate values of the mode, the median and the expected value. The values of these localization parameters differ very little for all three

Table 6. List of the aggregated lifetime location parameters for three criteria

| Parameter | Criterion <i>mo</i> | Criterion <i>me</i> | Criterion <i>ev</i> |
|-----------|---------------------|---------------------|---------------------|
| <i>mo</i> | 3894 | 3792 | 4080 |
| <i>me</i> | 3787 | 3688 | 3971 |
| <i>ev</i> | 3736 | 3637 | 3918 |

criteria, which confirms the previously noted large PDF symmetry of the obtained aggregated lifetimes $\tilde{T}_{mo}, \tilde{T}_{me}, \tilde{T}_{ev}$.

At the end of this article, the standard deviation (*sd*), the coefficient of variation (*cv*) and the skewness coefficient (*cs*) were calculated using the formulas (35), (36) and (37) for $T \sim wbl(\beta, \eta)$ and all three aggregated lifetimes:

$$sd(T) = \eta \left(\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right) \right)^{\frac{1}{2}} \quad (35)$$

$$cv(T) = \left(\frac{\Gamma\left(1 + \frac{2}{\beta}\right)}{\Gamma^2\left(1 + \frac{1}{\beta}\right)} - 1 \right)^{\frac{1}{2}} \quad (36)$$

$$cs(T) = \frac{ev(T) - mo(T)}{sd(T)} \quad (37)$$

The calculation results are summarized in Table 7.

The performed calculations show that considering the mode criterion, the standard deviation as well as the coefficients of variation and skewness have intermediate values compared to the other two criteria.

Table 7. List of lifetime characteristics for chosen criteria

| Characteristics | Criterion <i>mo</i> | Criterion <i>me</i> | Criterion <i>ev</i> |
|-----------------|---------------------|---------------------|---------------------|
| <i>sd</i> | 496 | 487 | 500 |
| <i>cv</i> | 0,1329 | 0,1339 | 0,1277 |
| <i>cs</i> | -0,3182 | -0,3168 | -0,3245 |

Moreover, as would be expected in all cases, the skewness is negative. In the presented example, the mode criterion was adopted as the result of the performed EEL procedure. The lifetime \tilde{T}_{mo} obtained according to this criterion has a Weibull distribution with a shape parameter of 8,998 and a scale parameter of 3945, i.e., $\tilde{T}_{mo} \sim wbl(8,998; 3945)$. For the obtained lifetime \tilde{T}_{mo} , graphs of the reliability function (Fig. 11) and the failure rate (Fig. 12) are prepared.

Potential lifetimes for selected failure probabilities, i.e., for $p = 0,01; 0,05; 0,1$ and $0,9$ are listed in the table 8. This informa-

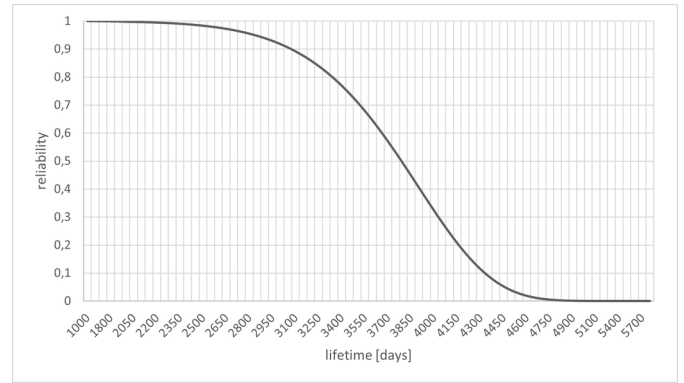


Fig. 11. Graph of the obtained reliability function

tion is very important in planning inspections of newly manufactured technical devices.

Using the formula (38), the failure rate function was determined (39) and then its graph was prepared (Fig. 12).

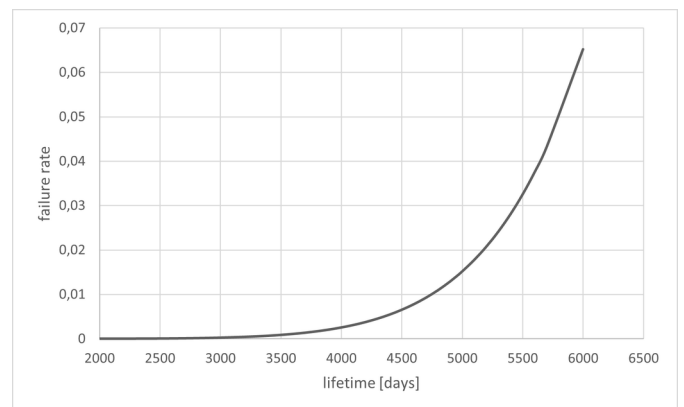
Table 8. List of the potential lifetimes

| <i>p</i> | 0,01 | 0,05 | 0,1 | 0,9 |
|--------------|------|------|------|------|
| t_p [days] | 2366 | 2836 | 3072 | 4328 |

$$\lambda_{wbl(\beta;\eta)}(t) = \frac{\beta}{\eta^\beta} t^{\beta-1} \quad (38)$$

$$\lambda_{wbl(8,998;3945)}(t) = (3,95289E - 32) t^{7,998}, t > 0 \quad (39)$$

Fig. 12. Graph of the predicted failure rate



Note that from the predicted failure rate obtained using the EEL procedure up to 4050 days of use of the TD in question, its failure rate will be less than 0,003. Of course, the final verification of the obtained results will take place in the process of using the tested technical devices.

7. Summary, conclusions, and orientations for future work

Maintenance of machinery and technical equipment under conditions of uncertainty requires the use of probabilistic prediction models in the form of lifetime distributions. Estimation of the parameters of these distributions is carried out with the use of statistical methods based on data about real life realizations of these objects. However, in cases when completely new solutions are introduced into exploitation,

we do not have such data and the only possible way of estimating the expected lifetime of these objects is the use of expert methods.

The paper proposes a modified method for estimating the probability distribution of the lifetime for new technical equipment based on expert assessments of parameters characterizing the potential lifetime of these objects. For the Weibull distribution, we use three parameters, two of which characterize the distribution and the third one to assess the quality of lifetime prediction by experts.

The innovation and originality of the developed lifetime parameter estimation procedure results from the application of this expert information for a specific lifetime model, instead of historical data. Such an approach to the issue of parameter evaluation has not yet been developed in the reliability theory.

The method is based on a subjective Bayesian approach to the problem of randomness and integrated with models of classical probability theory. A new procedure for expert elicitation of probabilities for any continuous random variable was developed, consisting of eight main steps. The first five steps have been developed based on good practices used in expert assessments of critical infrastructures. The sixth and seventh steps, which aim to objectivize the assessments of individual experts, are fully innovative. We propose that verification of the consistency of the expert judgements should be carried out using a consistency test. The results of this test can be used as a basis for assessing the credibility of the individual experts and for assigning appropriate weights to their opinions in the second part of this step. This will allow the quality of individual elicitation to be considered in the process of aggregating the opinions of different experts. The

aggregation process uses the ratings of all the experts, taking into account the weights estimated in the previous step, in order to obtain unambiguous data allowing the estimation of the parameters of the assumed probability distribution, which is a novelty not previously published in the literature.

Verification of the developed model on practical numerical examples for Weibull distribution has shown that the proposed method eliminates the basic limitations of the methods so far known and used in engineering practice. The calculations carried out demonstrated that considering the mode criterion, the standard deviation as well as the coefficients of variation and skewness have intermediate values compared to the other two criteria. Moreover, as would be expected in all cases, the skewness is negative. In the presented example, the mode criterion was adopted as the result of the performed Expert Elicitation of Lifetime procedure.

Further work of the authors will aim to generalize the developed method also to other probability distributions and to integrate this method with Bayesian inference process in operational decision making. This will require, among other things, consideration of economic aspects, and above all of the costs arising from the unreliability of the system under consideration.

Acknowledgement

The financial support for this research by the Rector's Grant No. 0213/SIGR/2154 of the Poznan University of Technology.

References

1. Abernethy RB. The New Weibull Handbook: Reliability & Statistical Analysis for Predicting Life. Safety, Survivability, Risk, Cost, and Warranty Claims (Fifth ed.), Florida, 2010.
2. Almalki SJ, Nadarajah S. Modifications of the Weibull distribution: A review. Reliability Engineering and System Safety 2014; 124: 32-55, <https://doi.org/10.1016/j.ress.2013.11.010>.
3. Andrzejczak K, Selech J. Quantile analysis of the operating costs of the public transport fleet. Transport Problems, 2017; 12 (3): 103- 111, <https://doi.org/10.20858/tp.2017.12.3.10>.
4. Aven T. Improving the foundation and practice of reliability engineering. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 2017, 231 (3): 295-305, <https://doi.org/10.1177/1748006X17699478>.
5. Beer M., Kougoumtzoglou IA, Patelli E. Emerging Concepts and Approaches for Efficient and Realistic Uncertainty Quantification. In: Frangopol DM, Tsompanakis Y. (eds.), Maintenance and Safety of Aging Infrastructure, 2014, Book Series "Structures & Infrastructures", Vol 10, Chapter 5, 121-154, CRC Press, Taylor & Francis Group, Boca Raton, London, New York, Leiden, <https://doi.org/10.1201/b17073-5>.
6. Biery, F., Hudak, D. and Gupta, S. Improving Cost Risk Analyses, Journal of Cost Analysis, Spring, 57-85, 1994, <https://doi.org/10.1080/08823871.1994.10462285>.
7. Book, S. A., Estimating Probable System Cost, Crosslink, 12-21, 2006.
8. Bourinet J M, Deheeger F, Lemaire M. Assessing small failure probabilities by combined subset simulation and support vector machines. Structural Safety 2011; 33(6): 343-353, <https://doi.org/10.1016/j.strusafe.2011.06.001>.
9. Bukowski L. Reliable, Secure and Resilient Logistics Networks. Delivering products in a risky environment. Springer Nature Switzerland AG: 2019, <https://doi.org/10.1007/978-3-030-00850-5>.
10. Carnevali L; Ridi L, Vicario E. A Quantitative Approach to Input Generation in Real-Time Testing of Stochastic Systems. IEEE Transactions on Software Engineering 2013. 39 (3): 292, <https://doi.org/10.1109/TSE.2012.42>.
11. Chaloner, K., Elicitation of Prior Distributions, in Berry, D.A. and Stangl, D.K. eds., Bayesian Biostatistics, New York: Marcel Dekker, 1996.
12. Galway, L.A. Subjective Probability Distribution Elicitation in Cost Risk Analysis, RAND Corporation, 2007, <https://doi.org/10.7249/TR410>.
13. Garvey, P.R. Probability Methods for Cost Uncertainty Analysis. 2000 New York: Marcel Dekker.
14. Grabski F. Semi-Markov Processes: Applications in System Reliability and Maintenance. 2014 Elsevier Inc., <https://doi.org/10.1016/B978-0-12-800518-7.00004-1>.
15. Hirose H. Bias correction for the maximum likelihood estimates in the two-parameter Weibull distribution. IEEE Transactions on Dielectrics and Electrical Insulation 1999; 6 (1): 66-68, <https://doi.org/10.1109/94.752011>.
16. <https://www.reliasoft.com/products/weibull-life-data-analysis-software>.
17. <https://www.statgraphics.com/life-data-analysis-and-reliability>.
18. <https://Wolfram Mathematica: Modern Technical Computing>.
19. <https://www.ttnews.com/articles/gauging-engines-life-expectancy-starts-b-life-rating>, 2016 June.
20. Jiang C, Zheng J, Han X. Probability-interval hybrid uncertainty analysis for structures with both aleatory and epistemic uncertainties: a review. Structural and Multidisciplinary Optimization 2018; 57(6): 2485-2502, <https://doi.org/10.1007/s00158-017-1864-4>.
21. Kahneman, D., Slovic, P. and Tversky, A. Judgment Under Uncertainty: Heuristics and Biases, Cambridge, UK: Cambridge University Press, 1982, <https://doi.org/10.1017/CBO9780511809477>.

22. Kaminskiy M, Krivtsov VV. A Simple Procedure for Bayesian Estimation of the Weibull Distribution. *IEEE Transactions on Reliability* 2005, 54 (4): 612-616, <https://doi.org/10.1109/TR.2005.858093>.
23. Khan M S, Pasha G R, Pasha A H, Reliability and Quantile Analysis of the Weibull Distribution. *Journal of Statistics* 2007; 14: 32-52.
24. Kozłowski E, Mazurkiewicz D, Kowalska B, Kowalski D. Application of multidimensional scaling method to identify the factors influencing on reliability of deep wells. In: Burduk A, Chlebus E, Nowakowski T, Tubis A. (eds) *Intelligent Systems in Production Engineering and Maintenance. ISPEM 2018. Advances in Intelligent Systems and Computing* 2019; 835: 56-65, https://doi.org/10.1007/978-3-319-97490-3_6.
25. Lacey P. An Application of Fault Tree Analysis to the Identification and Management of Risks in Government Funded Human Service Delivery. *Proceedings of the 2nd International Conference on Public Policy and Social Sciences* 2011. SSRN 2171117.
26. Meyer, M.A., and Booker, J.M. *Eliciting and Analyzing Expert Judgment: A Practical Guide*, Philadelphia, Pa.: Society for Industrial and Applied Mathematics and the American Statistical Association, 2001, <https://doi.org/10.1137/1.9780898718485>.
27. Morgan, M.G. and Henrion M., *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, New York: Cambridge University Press, 1990, <https://doi.org/10.1017/CBO9780511840609>.
28. Nobakhti A, Raissi S, Damghani K, Soltani R. Dynamic reliability assessment of a complex recovery system using fault tree, fuzzy inference and discrete event simulation. *Eksploracja i Niezawodność - Maintenance and Reliability* 2021; 23 (4): 593-604, <https://doi.org/10.17531/ein.2021.4.1>.
29. Pieniak D, Niewczas A M, Niewczas A, Bienias J. Analysis of Survival Probability and Reliability of the Tooth-composite Filling System. *Eksploracja i Niezawodność - Maintenance and Reliability* 2011; 2(50): 25-34.
30. Rinne H. *The Weibull Distribution: A Handbook*, 2008; CRC Press, New York, NY, <https://doi.org/10.1201/9781420087444>.
31. Selech J, Andrzejczak K. An aggregate criterion for selecting a distribution for times to failure of components of rail vehicles. *Eksploracja i Niezawodność - Maintenance and Reliability* 2020; 22 (1): 102-111, <https://doi.org/10.17531/ein.2020.1.1>.
32. Subjective Probability. *Best Practices in Dam and Levee Safety Risk Analysis*, 2019, <https://www.usbr.gov/ssle/damsafety/risk/BestPractices/Presentations/A6-subjectiveProbabilityPP.pdf>.
33. Sun B, Yang X, Ren Y, Wang Z, Antosz K, Loska A, Jasiulewicz-Kaczmarek M. Failure-based sealing reliability analysis considering dynamic interval and hybrid uncertainties. *Eksploracja i Niezawodność - Maintenance and Reliability* 2021; 23 (2): 278-284, <https://doi.org/10.17531/ein.2021.2.7>.
34. Wang J, Kalina M, Mesiar R, Jin L S. On some characteristics and related properties for OWF and RIM quantifier. *International Journal of Intelligent Systems* 2018; 33(6): 1283-1300, <https://doi.org/10.1002/int.21982>.
35. Wang Z. Method for Calculating the B10 Reliable Life of Mechanical Components of Vehicle Engine Based on the Stress-strength Interference. *Journal of Mechanical Engineering* 2014; 50(16): 47, <https://doi.org/10.3901/JME.2014.16.047>.
36. Wheeler, T.A., Hora, S.C., Cramond, W.R. and Unwin, S.D., *Analysis of Core Damage Frequency from Internal Events: Expert Judgment Elicitation*, Vol. 2, Washington, D.C.: Nuclear Regulatory Commission, NUREG/CR-4550, 1989.
37. Zaidi A, Bouamama B.O., Tagina M. Bayesian reliability models of Weibull systems: State of the art., *Int. J. Appl. Math. Comput. Sci.* 2012, 22 (3): 585-600, <https://doi.org/10.2478/v10006-012-0045-2>.
38. Zaman K, Rangavajhala S, McDonald M P, Mahadevan S. A probabilistic approach for representation of interval uncertainty. *Reliability Engineering & System Safety* 2011; 96: 117-130, <https://doi.org/10.1016/j.res.2010.07.012>.

A method for evaluating and upgrading systems with parallel structures with forced redundancy

Edward Michłowicz^a, Jerzy Wojciechowski^a

^aAGH University of Science and Technology, Faculty Mechanical Engineering and Robotics, al. Mickiewicza 30, 30-059 Kraków, Poland

Indexed by:




Highlights

- A system model with continuous delivery (24 hours a day) and forced oversupply is described.
- A method has been developed to assess the technical condition of the system.
- Indicators for assessing the status have been proposed.
- Exemplification for a complex underground primary mine drainage system

Abstract

The objects of the study are parallel-structure machine systems with redundancy associated with safety assurance of continuous material flow. The problem concerns systems in which the supply of materials takes place continuously (24 hours a day), and the system of operated machines must ensure the receipt and movement of the material at a strictly defined time and in the desired quantity. It is a system where the presence of a failure poses a threat to human life and environmental degradation. This paper presents a method for system condition assessment and upgrading for maintaining proper operation under conditions of continuous operation. A database of information about the current parameters of the system components (measurements, monitoring) is necessary for condition assessment. The method also uses lean techniques (including TPM). System evaluation and selection criteria for a suitable structure in terms of further operation were proposed. Exemplification was performed for an underground mine drainage system. As a part of the identification, selected parameters of the system components were measured, and their characteristics (motors, pumps, pipelines) were developed. The results of the analysis and the values of the adopted criteria were compared to the indicators for new pump sets. A two-option system upgrade was proposed, in addition to machine operating schedules, maintenance periods, and overhaul cycles.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

evaluation and retrofit method, parallel structure, redundancy, safety, process energy consumption.

1. Introduction

Maintaining a continuous flow of materials is one of the most important tasks in numerous operating systems. The problem is particularly relevant for systems with continuous operation, which additionally need to be resistant to hazardous environmental effects (safety function [23]). Security risk reduction through a security information transmission model was considered by Lei [17]. These systems should be resistant to abnormal disturbances (disasters) [5]. For a broad treatment of the safety assessment issue, see [16]. In case of tasks where failure of a work item results in mission failure, Levitin proposed models based on Poisson processes [18]. Redundancy of operational components is often used in the structures of such systems. A method for evaluating the security level of $k \times n$ systems was proposed by Młynarski [27], while a formulation using a Markov process for multi-state $k \times n$ systems was presented by Ruiz-Castro in his paper [29]. Balancing the probability of mission success and the risk of system failure by allocating redundancy has been described by Levitin in his publication [19]. Determining the optimal structure for these systems is the subject of numerous studies. A novel method for assess-

ing the reliability of multi-state systems based on structure learning algorithm was described by Li [21]. An optimal operation and maintenance schedule for $m \times n$ systems with reusable components was presented by Levitin in [20]. The use of Semi-Markov processes to assess readiness and reliability was demonstrated in the paper [31]. A reliability model for parallel systems under simultaneous failures was presented by Zhang [34]. In some areas, there are additional safety restrictions imposed by relevant directives and regulations. This is the case, for example, in aviation [10], or in offshore oil platform systems. Furthermore, the mining industry imposes additional restrictions on mine drainage (both in underground and open-pit mining). The problem with disposing of water of natural origin, flowing from the rock mass, is particularly important. The factors affecting the amount of inflowing water and the hazards resulting from them were described in many papers, such as by Bukowski [2] and others [7], [15], [24]. The effect of random factors on mine water inflow was analysed by Miladinović [26] and Quazizad [28]. For the safety of people and the operation of the deposits, the waters are pumped out using an appropriate system [3], [12], [33] and modern techniques [9], [13], [22]. These systems are very expensive to maintain and operate.

E-mail addresses: E. Michłowicz - michlowi@agh.edu.pl, J. Wojciechowski - jwojcie@agh.edu.pl

Having considered the foregoing, issues related to cost reduction are the focus of many papers, including those by Du Plessis [6], Gunson [9] and Afum [1], as well as Huang [14]. Over the last years, the issues of prevention and predictability were the focus of numerous researchers. A comprehensive approach to the issue was demonstrated in a paper by Werbińska [32]. Multi-criteria optimization for systems maintenance was proposed by Syan [30]. On the other hand, Han [11] proposed predictive strategies for multi-state systems, and Fauriat [8] proposed aperiodic control optimization based on information value. An example analysis of repair effectiveness using TPM techniques was presented in [4]. Most of the papers described herein deal with theoretical considerations related to typical reliability and operational problems. In the literature, there are no solutions related to forced redundancy. Cases of improper operation of such systems, known to the authors, were the genesis for the development of a method to evaluate and upgrade these systems.

2. General system model

An SPM material flow system is a certain ordered collection of E elements and R relationships between them (Figure 1):

$$\text{SPM} = \langle (E, R) \rangle = \langle \{X, Y, T\}, R \rangle,$$

$$T: X \Rightarrow Y$$

where:

$X = \{X_1, X_2, \dots, X_i, \dots, X_M\}$; for $i = 1, \dots, M$ – a set of external quantities describing the input elements (machines, material, among others),

$Y = \{Y_1, Y_2, \dots, Y_j, \dots, Y_N\}$; for $j = 1, \dots, N$ – a set of external quantities describing elements of the output (e.g., performance evaluation indicators, process performance),

$T = \{T_1, T_2, \dots, T_k, \dots, T_S\}$; for $k = 1, \dots, S$ – a set of quantities describing the transformation of the input vector processing into an output process,

$R = R_X \times R_Y \times R_T$ – material, information couplings between elements (X, Y, T) of the SPM system.

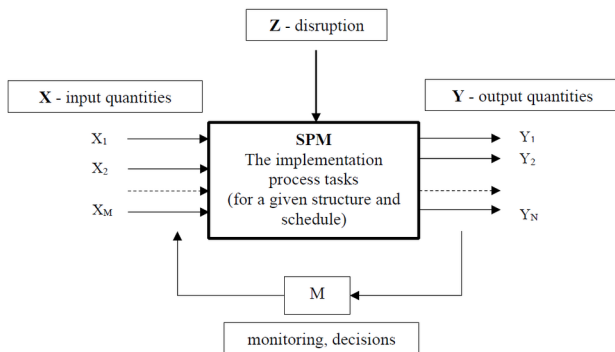


Fig. 1. SPM material flow system model

In the most general terms, two cases can be considered:

- developing new systems,
- modernization (retrofitting) of systems that have been in operation for many years.

In the studied system, the quantities that determine its specificity are (Figure 2):

- continuous supply of material (24 hours a day, all year round),
- the need to receive and move the material at the precise time and in the quantity requested,
- maintaining very high reliability of operation - system failure poses a threat to human life and can result in environmental degradation, hence the need to apply law-imposed safety conditions through a specific redundancy in the system structure.

Because of that, the quantities describing the outputs from the system should include information about the cost of process execution,

the efficiency achieved, the efficiency of operation, as well as the availability and utilization rate of the redundant system components.

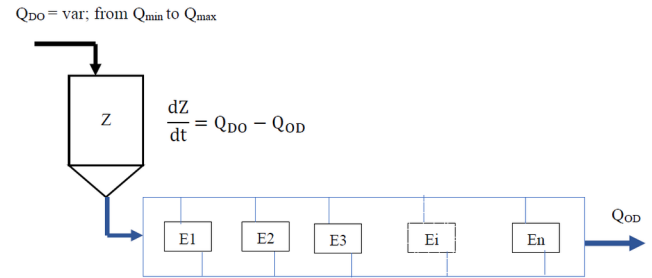


Fig. 2. Model of the continuous delivery and forced oversupply system

An example of redundancy forcing is shown in Figure 3.

For the case of 2-element operation ($n = 2$), the minimum number of system elements is $i = 2n + 1 = 5$.

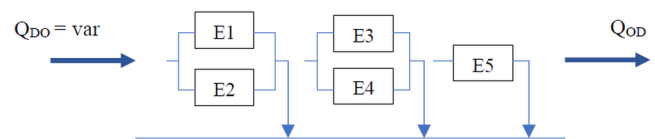


Fig. 3. A system model for the two-element operation case

In case of a three-element structure ($n = 3$), the minimum number of system elements is $i = 2n + 1 = 7$.

Observed cases of improper operation of these systems became the genesis for the development of a method to assess their technical condition. The assessment result is a variant solution – continue operating the system or upgrade it.

3. The method of assessing the condition and upgrading the system

The proposed method is multi-stage (consisting of seven stages). A simplified block diagram of the method is shown in Figure 4.

The main components of each stage are described below.

Stage I – Process identification – Actions: 1-2-3-4-5-6

1. Selecting a process for analysis.
2. Establishing security constraints for system operation (directives, industry regulations).
3. Drawing up an accurate process diagram (process structure, including forced redundancy of components).
4. Identifying the basic quantities that describe the process.
5. Determining the parameters (characteristics) describing the assumed quantities (making measurements, necessary calculations).
6. Collection of process data (database, including historical).

Stage II – Identifying machine functioning – actions: 7-8-9-10

1. Describing losses and waste in the process (e.g., 7 muda, 6 big losses).
2. Identifying machine downtime and damage.
3. Drawing up a Pareto diagram – causes of downtime. Selecting causes for improvement.
4. Setting targets – MTTR and MTBF limits.
5. Determination of OEE effectiveness measure.

Stage III – Establishing criteria for evaluating system condition and performance – actions: 12-13

1. Defining criteria for evaluating the system.
2. Determining the values of evaluation indicators (measurements, calculations).

Stage IV – Analysing system effectiveness – actions: 14 – 16

1. Analysing the compliance of the identified redundant structure with process requirements and enforced constraints.
2. Analysing the timing of system operations.
3. Analysing the system performance evaluation metrics obtained.

Stage V – Evaluating the system and selecting a strategy for further action – actions 17-18

1. If the assessment complies with the adopted criteria – further operation in accordance with the implemented schedule.
2. If the assessment is non-compliant – a proposal to upgrade the system.

Stage VI – Implementing changes – actions: 20-21

1. Making changes to improve system performance evaluation metrics.
2. Developing a schedule for machine operation, maintenance, and overhaul.
3. Developing a schedule (implementation map) for system upgrades.

Stage VII - Analysing the effects and improving - actions: 22-23

1. Analysing effects after making changes.
2. Persistent implementation of kaizen principles!

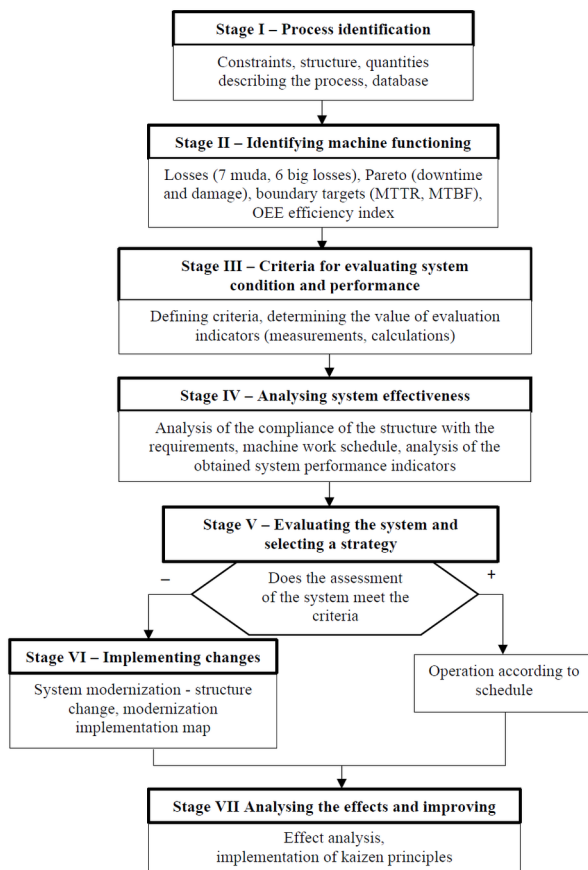


Fig. 4. Block diagram of the system state assessment method

4. Exemplification – the main drainage system

4.1. System identification

The main water drainage system analysed (Figure 5), is located in an underground mine at the 500 m level. Ten pumping units (P1 to P10) consisting of OW250/8 pumps and SCUd134u motors are installed in the main drainage pumping station. Each pumping unit is connected to two pressure pipelines with diameters of 500 mm,

through which water is pumped to the surface at the height of $H = 500$ meters. Based on the hydrological conditions and the size of the underground excavations, the projected water supply is $0.28 \text{ m}^3/\text{sec}$ ($16.83 \text{ m}^3/\text{min}$), which means that the daily water supply is equal to 24235 m^3 . The capacity of the water roads in which water is collected is 20196 m^3 . The requirements for the main drainage equipment are governed by the *Regulations of the Minister of Energy of 2016* and stipulate that the discharge of the daily inflow of water must be realized in a time not exceeding 20 hours, and the minimum number of pumps is determined by the relation: $i = 2n + 1$ (n – the calculated number of pumps).

A schematic of the main drainage system under study is shown in Figure 5.

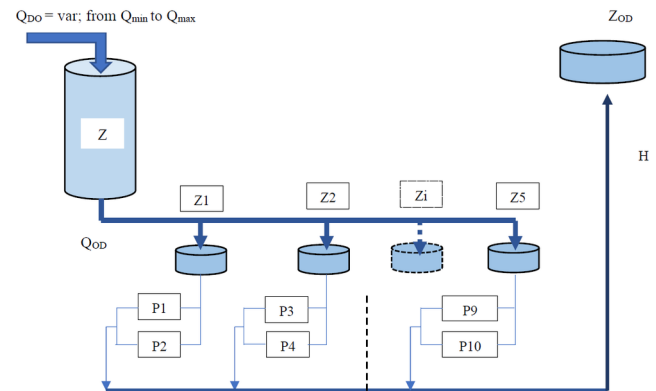


Fig. 5. Diagram of the main drainage system

With the forecasted mine water inflow, the required total pumping capacity, meeting the limitations of the mining regulations, is $Q = 20,20 \text{ m}^3/\text{min}$ (pumping for 20 hours), while the pumping head $H_u = 530 \text{ m}$. The eight-stage OW250/8 pumps installed in the pumping station have a rated capacity of $Q = 8.33 \text{ m}^3/\text{min}$ ($500 \text{ m}^3/\text{h}$) and a head $H_u = 560 \text{ m}$. The requirements specified by mining regulations are met with two pumps working continuously and the third pump working half the time. Having considered that, the number of pumps required (assuming a total number of pumps $n=3$) is seven – according to the rules: $i = 2n + 1$. There are ten pumping units in the pumping station, i.e., the main drainage system analysed is definitely overdimensioned.

The basic principles of monitoring the technical condition of pumping units were described by Nowicki [21, 22], and other diagnostic tests related to the test object (drainage system) were presented in papers [25] and [35].

A complete analysis of the main drainage system operation included:

1. Measurement of operating parameters
 - analysis of water composition and quality,
 - determination of pump characteristics,
 - characteristics of flow pipelines (manifolds),
 - the suction height of the pumping system.
2. Assessing the technical condition of pump units
 - testing the thickness of the discharge manifold walls,
 - measuring the power consumed by the pump,
 - vibro-acoustic diagnostics of pump units.
3. Qualitative and quantitative assessment and classification of failures and damages for a 5-year period.

To calculate the parameters of the flow characteristics of the pumps, known relationships were used to determine the values:

- useful lifting height H_u ,
- c velocities of water in the suction and discharge ports,
- P_u power output transferred to the pumped water flow,
- η_{zp} efficiency of the pump unit (related to the power of electric motors),

- P_m power on the pump shaft,
- η_p pump efficiency.

In order to evaluate the condition of the pumps, sections of the catalogue characteristics in terms of measured changes in pump performance are plotted on the figures. Figures 6 and 7 show examples of the characteristics of pump #1 (P1).

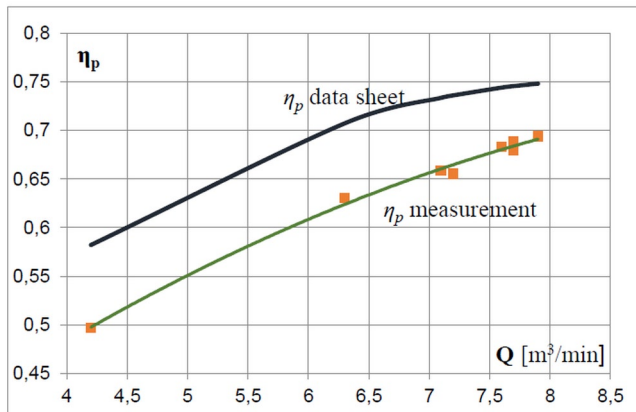


Fig. 6. P1 pump utility power characteristic

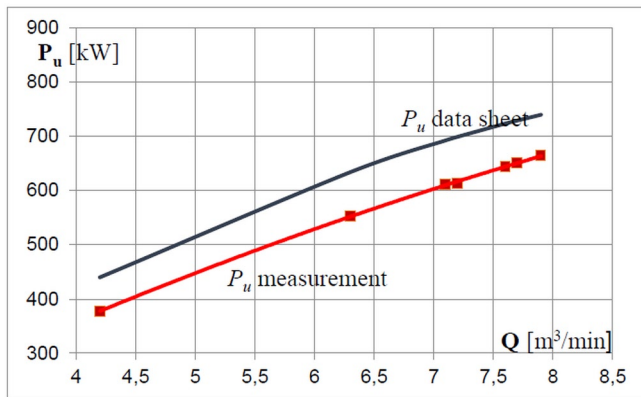


Fig. 7. Efficiency characteristics of pump P1

The efficiency of pump no. 1 is lower than catalogue efficiency from about 8% at 8 m³/min to 14% at 4.5 m³/min. The nature of pump operation means that the useful head characteristics are minimally affected by throttling. The decisive factor is the geometric head $H_g = 489$. The pressure increase in the pump depending on the flow resistance is relatively small, amounting to a few percent (approx. 5% on average) with respect to the geometric head.

Table 1. Comparison of operating parameters of pump units at maximum efficiency

| | | P 1 | P 2 | P 3 | P 4 | P 5 | P 6 | P 7 | P 8 | P 9 | P 10 | average |
|---|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|
| Volumetric stream Q | m³/min | 7,9 | 8,1 | 9,2 | 8,1 | 8,5 | 8,4 | 8,0 | 9,2 | 7,6 | 8,2 | 8,3 |
| Effective head H_u | m | 514,7 | 523,6 | 535,1 | 508,8 | 516,4 | 513,9 | 498,2 | 514,0 | 503,3 | 498,2 | 512,6 |
| Electrical power P_{el} | kW | 1030 | 1214 | 1200 | 1195 | 1263 | 1275 | 1266 | 1399 | 1137 | 1314 | 1229 |
| Pump efficiency η_p | -- | 0,694 | 0,614 | 0,721 | 0,615 | 0,613 | 0,595 | 0,553 | 0,594 | 0,591 | 0,547 | 0,614 |
| Relative pump efficiency η_p/η_{pk} | -- | 0,93 | 0,82 | 0,97 | 0,82 | 0,81 | 0,79 | 0,74 | 0,80 | 0,79 | 0,73 | 0,82 |
| Energy consumption q_p | kWh/m³ | 2,173 | 2,498 | 2,174 | 2,459 | 2,476 | 2,530 | 2,638 | 2,534 | 2,493 | 2,671 | 2,465 |
| Total pump operating time $\Delta\tau$ | h | 3336 | 12032 | 757 | 21000 | 24000 | 29017 | 24015 | 6015 | 23473 | 32041 | 17569 |
| Pump operating time $\Delta\tau_R$ | h | 238 | 3142 | 1667 | 1470 | 1857 | 646 | 299 | 2883 | 1936 | 3837 | 1634 |

Pi – pump units, $i = 1 \div 10$.

Similar characteristics were developed for all other pumps (P2 through P10). Additionally, characteristics were developed for each pump unit:

- flow Q in relation to the discharge height H ,
- power output P as a function of stream flow rate Q .

4.2. Analysis of the study results

Efficiency, energy consumption, and unit pumping costs were determined for all of the main drainage pumping units studied. The results obtained are placed in Table 1.

The value of the quotient of the efficiency ratio η_p and the catalogue efficiency η_{pk} of the pump at a fixed water flow was taken as a quality measure of the pump condition. A smaller quotient value indicates a worse condition of the operating pump. For the pumps tested, the value of the η_p/η_{pk} quotient takes values in a wide range. For pumps 1 and 3, it has a value above 0.90, while for pump 10, it is only 0.73. The condition of pumps for which this quotient takes values below 0.80 should be considered unsatisfactory. The average value for all pumps in the pumping station is $\eta_p/\eta_{pk} = 0.82$ (relatively low, close to unsatisfactory).

Table 1 also shows the coefficients determining the pumps energy consumption q_p . The q_p coefficient determines the amount of electricity in kWh needed to pump out 1 m³ of water. This is one of the most important indicators of system evaluation, as it directly affects operating costs. For the pumping units of the studied pumping station, the energy consumption of the water pumping-out process takes the values $q_p = 2.173 \div 2.671$ kWh/m³, whereas the mean value is $q_p = 2.465$ kWh/m³. The value of costs should be related to the current price per unit of delivered electricity. The energy consumption of water pumping is shown in Figure 8, with the red line indicating the average value for the main drainage pumping stations.

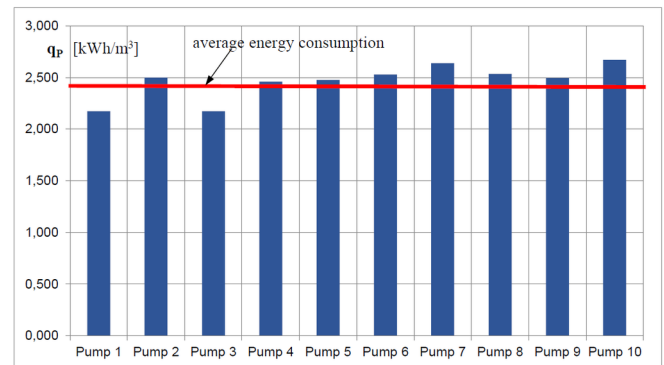


Fig. 8. Energy consumption of the pumping-out process

Figure 9 shows the efficiency of the pumps as a function of operating time – it can be observed that operating time above 5,000 hours

results in a clear drop in efficiency. The dependence of energy consumption on operating time is illustrated in Figure 10. The course of the energy consumption curve is obviously the opposite of the efficiency characteristics.

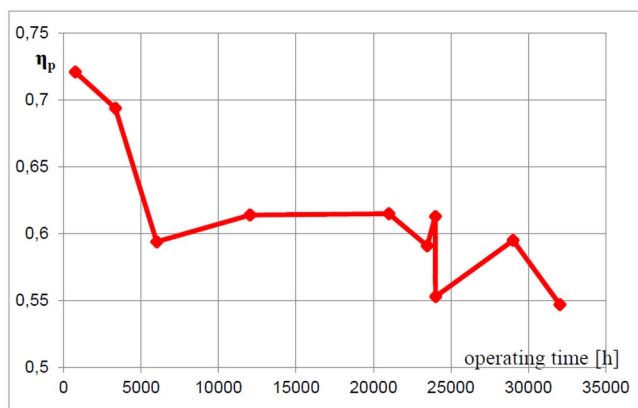


Fig. 9. Dependence of average pump efficiency on time

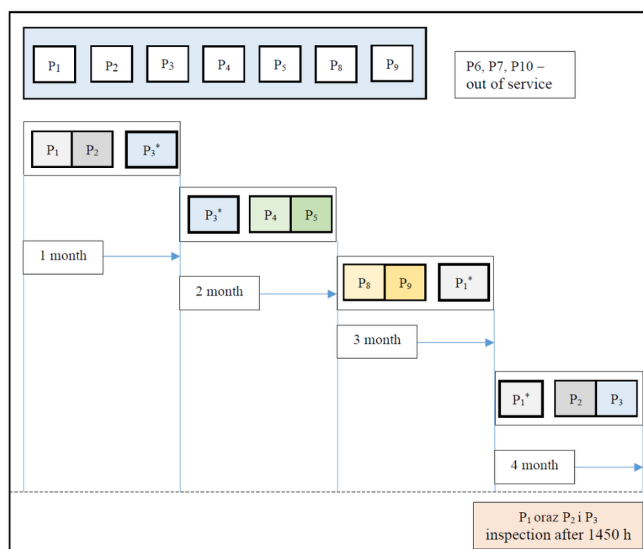


Fig. 10. Time dependence on the mean energy consumption

Above 5000 hours, there is a deterioration of pump technical condition due to operational wear, which results in an increase in demand for electricity to pump out 1 m^3 of water – from about 2.17 to 2.67 kWh/ m^3 .

4.3. Suggested changes

The conducted study and the analysis of the obtained results allow to clearly state that the assessment of the drainage system technical condition is unsatisfactory. According to the proposed evaluation method (stages V and VI – Figure 4), a system upgrade is required. The most significant elements of the proposed upgrade are:

- taking out of service units showing high wear (P6, P7, P10),
- double-variant operational improvement (for existing and new units),
- developing a model schedule for units (Figure 11),
- developing an implementation map for system upgrades (several years, high purchase and investment costs).

OPTION 1 – for existing units (motor + pump)

Proposed model unit operation schedule (3 + 3 + 1).

Decommissioning of units: P6, P7, P10, monthly unit cycles.

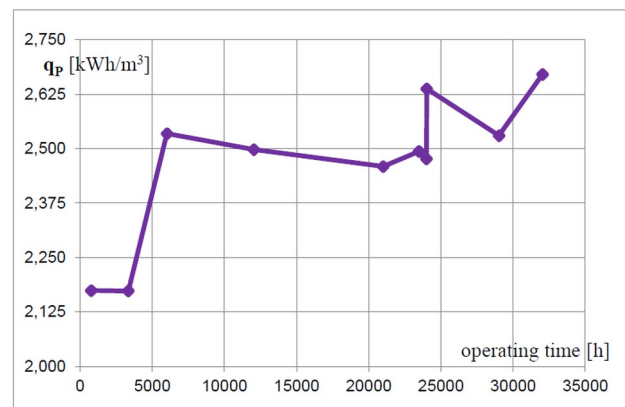


Fig. 11. Work schedule for pumping units – variant (3+3+1)

OPTION 2 – for new units (purchase)

Structure: 5 pumping units; arrangement (2 + 2 + 1).

unit selection (motor + pump): to be determined – the best.

Selection criteria: performance, efficiency, electrical power, price.

5. Summary

In systems executing tasks associated with continuous supply of the material (24 hours a day) and with limitations imposed on the reception and movement of these materials, there is a necessity to apply redundancy of the system elements. In case of a task in which the failure of the system poses a threat to human life and leads to environmental degradation, there are additional rules applied to determine the necessary redundancy (dependence on the industry, industry branch). The encountered cases of improper operation of these systems were the basis for the development of an original method of condition assessment and retrofitting to improve system evaluation indices. The proposed method is multi-stage (consisting of seven stages) and requires numerous identification tests, measurements, and calculations. However, it does result in correct operation of the system and a clear reduction in costs.

The application of the developed method is presented on the example of the main drainage pumping station located at 500 m level of an underground mine. The system consists of ten pumping units (P1 to P10). Continuous pumping activity requires the operation of 2.5 pumping units with a total capacity of $20.20\text{ m}^3/\text{min}$. This results in a requirement for 7 pumping units (as per $i = 3+3+1$). The analysed system is thus clearly overdimensioned. The measurement results and the characteristics and quality indicators determined from them demonstrate unsatisfactory or poor condition of most pumps. This is mainly a consequence of long pump operation times with no overhauls (with 6 pumps working for over 20,000 hours). Analyses show that the operation runs properly up to 5,000 operating hours, with a clear drop in efficiency above this number. Most pumps are about 20% less efficient than the catalogue efficiency of new pumps. Pumping efficiencies as low as those obviously translate into increased energy consumption and unit cost of pumping the water out (with energy consumption increasing from 2.173 to 2.671 kWh/ m^3). It can be concluded that the energy consumption and the cost of the pump-out process increases at the same rate as the efficiency decreases, which is about 20%. System upgrades are required due to the high energy consumption and operating cost ratios. Modernization should include the gradual installation of new pumps, with significantly better technical and economic indicator values. The authors proposed a new solution for this system, in which one can optionally choose a version with five (2+2+1) new units (more expensive solution) or with seven (3+3+1) pump units. A schedule for implementing changes to the system was also proposed as a part of the modernization.

References

1. Afum B.O, Ben-Awuah E. A Review of Models and Algorithms for Surface-Underground Mining Options and Transitions Optimization: Some Lessons Learned and the Way Forward. *Mining* 2021; 1(1): 112-134, <https://doi.org/10.3390/mining1010008>.
2. Bukowski P. Evaluation of water hazard in hard coal mines in changing conditions of functioning of mining industry in Upper Silesian Coal Basin - USCB (Poland). *Archives of Mining Sciences* 2015; 60(2): 455-475, <https://doi.org/10.1515/amsc-2015-0030>.
3. Chen T, Riley C, Van Hentenryck P, Guikema S. Optimizing inspection routes in pipeline networks. *Reliability Engineering & System Safety* 2020; 195: 106700, <https://doi.org/10.1016/j.res.2019.106700>.
4. Daniewski K, Kosicka E, Mazurkiewicz D. Analysis of the correctness of determination of the effectiveness of maintenance service actions. *Management and Production Engineering Review* 2018; 9(2): 20-25.
5. Dudek D, Nowakowski T. Resilience engineering - agents of open pit mining machine disasters in Poland. In: *Mining machines and earth-moving equipment : problems of design, research and maintenance / Marek Sokolski ed.* Cham : Springer 2020; 1-20, https://doi.org/10.1007/978-3-030-25478-0_1.
6. Du Plessis GE, Arndt DC, Mathews EH. The development and integrated simulation of a variable flow energy saving strategy for deep mine cooling systems. *Sustainable Energy Technologies and Assessments* 2015; 10: 71-78, <https://doi.org/10.1016/j.seta.2015.03.002>.
7. Fan L, Ma X. A review on investigation of water-preserved coal mining in western China. *International Journal of Coal Science & Technology* 2018; 5: 411-416, <https://doi.org/10.1007/s40789-018-0223-4>.
8. Fauriat W, Zio E. Optimization of an aperiodic sequential inspection and condition-based maintenance policy driven by value of information. *Reliability Engineering & System Safety* 2020; 204: 107133, <https://doi.org/10.1016/j.res.2020.107133>.
9. Gunson AJ, Klein B, Veiga M, Dunbar S. Reducing mine water network energy requirements. *Journal of Cleaner Production* 2010; 18(13): 1328-1338, <https://doi.org/10.1016/j.jclepro.2010.04.002>.
10. Gołda P, Zawisza T, Izdebski M. Evaluation of efficiency and reliability of airport processes using simulation tools. *Eksploracja i Niezawodność - Maintenance and Reliability* 2021; 23 (4): 659-669, <https://doi.org/10.17531/ein.2021.4.8>.
11. Han X, Wang Z, Xie M et al. Remaining useful life prediction and predictive maintenance strategies for multi-state manufacturing systems considering functional dependence. *Reliability Engineering & System Safety* 2021; 210: 107560, <https://doi.org/10.1016/j.res.2021.107560>.
12. Hancock S, Wolkersdorfer C. Renewed demands for mine water management. *Mine Water Environ* 2012; 31(2): 147-158, <https://doi.org/10.1007/s10230-012-0176-6>.
13. Hu L, Zhang M, Yang Z, Fan Y. Estimating dewatering in an underground mine by using a 3D finite element model. *PLOS ONE* 2020; 15(10): e0239682.
14. Huang S, Li G, Ben-Awuah E, Afum BO, Hu N. A stochastic mixed integer programming framework for underground mining production scheduling optimization considering grade uncertainty. *IEEE Access* 2020; 8: 24495-24505, <https://doi.org/10.1109/ACCESS.2020.2970480>.
15. Inung A, Adnyano A, Bagaskoro M. Technical study of mine dewatering system in coal mining PROMINE 2020; 1: 28-33, <https://doi.org/10.33019/promine.v8i1.1794>.
16. Jemai H, Badri A, Ben Fredj N. State of the Art and Challenges for Occupational Health and Safety Performance Evaluation Tools. *Safety* 2021; 7(3): 64, <https://doi.org/10.3390/safety7030064>.
17. Lei Y, Wu C, Feng Y, Wang B. Optimization of multi-level safety information cognition (SIC): A new approach to reducing the systematic safety risk. *Reliability Engineering & System Safety* 2019; 190: 106523, <https://doi.org/10.1016/j.res.2019.106497>.
18. Levitin G, Finkelstein M, Dai Y. Optimal preventive replacement policy for homogeneous cold standby systems with reusable elements. *Reliability Engineering & System Safety* 2020; 204: 107135, <https://doi.org/10.1016/j.res.2020.107135>.
19. Levitin G, Finkelstein M, Li Y. Balancing mission success probability and risk of system loss by allocating redundancy in systems operating with a rescue option. *Reliability Engineering & System Safety* 2020; 195: 106694, <https://doi.org/10.1016/j.res.2019.106694>.
20. Levitin G, Xing L, Dai Y. Optimal operation and maintenance scheduling in m-out-n standby systems with reusable elements. *Reliability Engineering & System Safety* 2021; 211: 107582, <https://doi.org/10.1016/j.res.2021.107582>.
21. Li J, Wang Z, Ren Y, Yang D, Lv X. A novel reliability estimation method of multi-state system based on structure learning algorithm. *Eksploracja i Niezawodność - Maintenance and Reliability* 2020; 22 (1): 170-178, <https://doi.org/10.17531/ein.2020.1.20>.
22. Liao M, Si Q, Fan M, Wang P, Liu Z, Yuan S, Cui Q, Bois G. Experimental Study on Flow Behavior of Unshrouded Impeller Centrifugal Pumps under Inlet Air Entrainment Condition. *International Journal of Turbomachinery, Propulsion and Power* 2021; 6(3): 31, <https://doi.org/10.3390/ijtp6030031>.
23. Mancuso A, Compare M, Salo A, Zio E. Portfolio optimization of safety measures for the prevention of time-dependent accident scenarios. *Reliability Engineering & System Safety* 2019; 190: 106500, <https://doi.org/10.1016/j.res.2019.106500>.
24. Masood N, Hudson-Edwards K, Farooqi A. True cost of coal: coal mining industry and its associated environmental impacts on water resource development. *Journal of Sustainable Mining* 2020; 19(3): 1, <https://doi.org/10.46873/2300-3960.1012>.
25. Menegaki M, Damigos D. A systematic review of the use of environmental economics in the mining industry. *Journal of Sustainable Mining* 2020; 19(4): 254-271, <https://doi.org/10.46873/2300-3960.1034>.
26. Miladinović B, Vakanjac V, Bukumirović D, Dragišić V. Simulation of Mine Water Inflow: Case Study of the Štavalj Coal Mine (Southwestern Serbia). *Archives of Mining Sciences* 2015; 60(4): 955-969, <https://doi.org/10.1515/amsc-2015-0063>.
27. Młynarski S, Pilch R, Smolnik M, Szybka J, Wiązania G. A Method for rapid evaluation of k-out-of-n systems reliability. *Eksploracja i Niezawodność - Maintenance and Reliability* 2019; 21 (1): 170-176, <https://doi.org/10.17531/ein.2019.1.20>.
28. Qazizada ME, Pivarčiová E. Reliability of parallel and serial centrifugal pumps for dewatering in mining process. *Acta Montanistica Slovaca* 2018; 23(2): 141-152.
29. Ruiz-Castro JE. A complex multi-state k-out-of-n: G system with preventive maintenance and loss of units. *Reliability Engineering & System Safety* 2020; 197: 106797, <https://doi.org/10.1016/j.res.2020.106797>.
30. Syan C, Ramsoobag G. Maintenance applications of multi-criteria optimization: A review. *Reliability Engineering & System Safety* 2019; 190: 106520, <https://doi.org/10.1016/j.res.2019.106520>.
31. Świderski A, Borucka A, Grzelak M, Gil L. Evaluation of Machinery Readiness Using Semi-Markov Processes. *Applied Sciences* 2020;

- 10(4): 1541, <https://doi.org/10.3390/app10041541>.
32. Werbińska-Wojciechowska S. Preventive Maintenance Models for Technical Systems. In: Technical System Maintenance: Delay-Time-Based Modelling. Cham: Springer International Publishing 2019, <https://doi.org/10.1007/978-3-030-10788-8>.
 33. Tang Y, Zheng G, Zhang S. Optimal control approaches of pumping stations to achieve energy efficiency and load shifting. *Electrical Power and Energy Systems* 2014; 55: 572-580, <https://doi.org/10.1016/j.ijepes.2013.10.023>.
 34. Zhang C, Zhang Y. Common cause and load-sharing failures-based reliability analysis for parallel systems. *Eksploracja i Niezawodność - Maintenance and Reliability* 2020; 22 (1): 26-34, <https://doi.org/10.17531/ein.2020.1.4>.

Integrating advanced measurement and signal processing for reliability decision-making

Indexed by:



Edward Kozłowski^a, Katarzyna Antosz^b, Dariusz Mazurkiewicz^c, Jarosław Sęp^b, Tomasz Żabiński^d

^aLublin University of Technology, Faculty of Management, ul. Nadbystrzycka 38, 20-618 Lublin, Poland

^bRzeszów University of Technology, Faculty of Mechanical Engineering and Aeronautics, ul. Powstańców Warszawy 8, 35-959, Rzeszów, Poland

^cLublin University of Technology, Mechanical Engineering Faculty, ul. Nadbystrzycka 36, 20-618 Lublin, Poland

^dRzeszów University of Technology, Faculty of Electrical and Computer Engineering, ul. W. Pola 2, 35-959 Rzeszów, Poland


Highlights

- Force and torque sensors analysed as an alternative to the vibration measurement.
- Effective condition prediction when integrated with adequate signal processing.
- Decision trees with various types of wavelets selected for predictive models.
- High accuracy method proposed to trace tool condition in real-time.

Abstract

An advanced milling machine multi-sensor measurement system as a condition monitoring tool was presented. It was assumed that the data collected from the 3-axis force and torque sensor can be used as a new approach and an alternative to the typical vibration signal based health monitoring and remaining useful life prediction (RUL), when integrated with machine learning techniques that are regarded as a powerful solution. Measurement system integration with the proposed signal processing method based on decision trees with different types and levels of wavelets for the cutter reliability decision-making process was presented together with proving their ability to trace the tool condition accurately. Prediction errors achieved with the use of different signal sources and data processing methods were presented and compared.

Keywords

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>) 

force and torques measurement, condition monitoring, cutting tool, remaining useful life, prediction.

1. Introduction

Innovative technological machines are constructed as advanced mechatronic systems facing extremely high demands with respect to their performance, reliability and product quality. In both, their construction and operation, the problem of maximum productivity, where several factors such as efficiency, production costs or resources and energy consumption must be taken into account, is important too, all in the context of sustainable manufacturing requirements [21, 32]. Machine tools together with other technological machines used in production systems of high technology industry form complex systems functioning as Industry 4.0 elements. Such terms as Industrial Internet of Things (IIoT) or Machine to Machine Communication (M2M) do not only describe the current industrial revolution but they also characterize any modern machine tool. According to the paradigm of the fourth industrial revolution complemented by the mentioned high demands in quality and reliability, machine tools are equipped with several sensors, diagnostic and monitoring systems.

The measuring methods of machine tool key elements wear are classified as direct (intermittent, offline) and indirect [28, 45, 49]. For example, tool wear is measured based on various sensor signals containing cutting force, torque, vibration, acoustic emission, sound, surface roughness, temperature, displacement or spindle power. The features of the signals correlating to the tool wear are captured to monitor

tool condition and to do this, a mass of signal processing methods were used, such as time series modeling, Fast Fourier Transform and time–frequency analysis, the amount of data gathered and calculation involved in corresponding parameters with tool wear is enormous. According to the detailed analysis presented in [49], up to now, many types of sensors and signal processing techniques are used in machine tool and especially in cutting tool condition monitoring and RUL prediction. However, most of these sensors are wired, mounted inconveniently on the machine during the machining operations, and the prognostic information is not easy to be integrated into the manufacturing system [28, 31, 45, 49]. One of the problems is huge amount of data gathered. As a result, we face two types of research challenges concerning machine tools systems which are actually strongly interconnected. First of all, we force the problems related to their adequate measurement techniques for service life, health monitoring and reliability, especially with respect to predicting future states in order to enable the inference and implementation of executive activities in terms of failure-preventing servicing [24]. In addition, potential new solutions according to digital era requirements have to go beyond typical tool wear monitoring methods in real-time by tracking for example force model coefficients during the cutting process [30, 32]. On the other hand, diagnostic or maintenance systems require an operator to make reliable predictions and decisions under uncertainty. All these

E-mail addresses: E. Kozłowski - e.kozlovski@pollub.pl, K. Antosz - katarzyna.antosz@prz.edu.pl, D. Mazurkiewicz - d.mazurkiewicz@pollub.pl, J. Sęp - jsztmiop@prz.edu.pl, T. Żabiński - tomz@prz.edu.pl

aspects create the so called information overload problem, which can be solved with the use of data mining and existing data reduction techniques. Unfortunately, in complex production systems machinery operating under diverse conditions requires more advanced measurement and data processing approaches. As it was pointed by Zhao et al. [50], as a key component in modern manufacturing system, machine health monitoring has fully embraced the big data revolution. In order to extract useful knowledge, to create information based on it and, finally, to make appropriate decisions from the big data, machine learning techniques were regarded as a powerful solution. Machine learning has a potential for improving products and processes, enabling successful predictions using past experience, data and information. It encompasses several algorithms and tools used for a vast array of different data processing tasks [6, 40]. However, as mentioned by Ahamad et al. [1], the Big Data analytics require innovative tools that address the challenges faced by data volume, variety and velocity. Especially, when data fusion - integration of data and knowledge from several sources is necessary to be taken into consideration [23, 48]. These techniques may act as an effective bridge connecting advanced machinery sensors systems, big data and intelligent machine health monitoring systems. On one hand, it requires monitoring systems equipped with adequate sensors in order to collect the data. On the other hand, the transition from raw industrial big data to knowledge-based executive actions without any human action also requires the development of new analytical tools. This means another need of new expert and intelligent systems. For the purpose of mechanical systems development, studies must be conducted particularly on the measurement systems construction, and further the development of the integrated analytical solutions for intelligent modules that take advantage of a data analysis and intelligent decision support tools in order to predict and prevent a potential failure of machines or their crucial elements [24]. Machine tools are considered as a representative example of such studies needs and their real-world applications. For example, the new methods enable early prediction of the machine tool remaining useful life, its current condition classification, or both of them simultaneously.

For this purpose, numerous research works have been carried out providing new knowledge, although not without several weaknesses that should be solved. Condition monitoring techniques such as temperature, vibration or acoustic signal analysis, play an important role as indicators of a developmental failure, and have a wide range of different applications for the purpose of fault diagnosis. These different applications in technological machines such as machine tools or in any other mechanical systems, allow to compare the proposed methods and achieved results. Vamsi et al. [43] simulated the non-stationary load profile acting on a wind turbine. The vibration, acoustic signal and lubricating oil signals were simultaneously acquired. The raw signals were processed using a wavelet-based feature extraction technique. Next, the efficiency of each of these condition monitoring techniques under stationary and non-stationary loads were compared by using Support Vector Machine (SVM) as the classification technique. A decision tree algorithm was used to identify among the extracted features the dominant one (which was a standard error). SVM was used to classify the features among the fault levels. The main purpose of these investigations was to verify diagnostic capabilities of vibration signal analysis, compared to other techniques in the fault detection of a gear tooth root crack and a gear tooth chip. The authors [2] did not go beyond the standard diagnostic. The RUL prediction with the use of the collected data and modelling techniques as a decision support tool, unfortunately were not taken into consideration. The remaining useful life prediction via the combined use of the SVM as a classification tool and AutoRegressive and Integrated Moving Average (ARIMA) based identification as an expert system tool for the real-time monitoring of a manufacturing process was presented by Kozłowski et al. [24]. The objective of the study was to develop a new method for the proper estimation and representation of uncertainty in the RUL prediction. Therefore, in the analysed case, sensor

data management involved the application of the SVM in order to construct a classifier for the cutter condition assessment, investigation of the effect of an acoustic signal correlation displacement length on the diagnostic error and the number of support vectors, and, finally, the development of the RUL prediction method. Several different advances in modelling of metal machining processes were also analysed in details by Arrazola et al. [2], as a result of a significant progress in developing industry-driven predictive models. The authors claim that the operation-level predictive models still need to be developed, especially for direct, industrial applications. As a part of a similar research project, mechanics of the milling system with serrated end-mills were studied by Pelayo et al [34], using force and surface topography models. A stationary milling force model was developed to predict the resulting machined surfaces. The authors point that the available cutting tools in standard catalogues are not homogeneous from one seller to another. Large differences are seen on the tool's features suggesting that there is not a unified criterion. It requires more effective measurement systems but also more universal analytical tools. The spindle bearing system as one of the most important parts of a machine tool was the subject of dynamic modelling by Xi et al. [47]. Based on the developed spindle bearing system model, the dynamic response of the system with different cutters and under different cutting conditions was simulated and compared with the experiment measured results. The presented results show that the simulated responses are in accordance with the experiment measured responses. However, they were achieved only in the laboratory conditions that do not directly reflect actual industrial production. According to [22], when avoiding chatter and improving machining efficiency and accuracy, the machining process analysis is extremely important. This type of analysis is essential in order to enable high productivity without sacrificing surface quality and inducing significant surface errors. Its exact implementation depends on the dynamics modelling with a reliable requirement of the system's dynamic parameters. With regard to this, a novel model testing strategy was proposed for obtaining the system's dynamic parameters. A triaxial acceleration sensor was used there and the related parameter processing techniques were proposed. Unfortunately, for validation, only two cutters with different diameters were employed in the experiments what makes the achieved results very limited.

As proved by Bousdekis et al. [4], the emergence of Industry 4.0 led to a wide use of sensors which facilitate manufacturing operations. Machining technology is one of the core examples. Predictive maintenance has significantly benefited from these technological advancements with the use of real-time detection and prediction algorithms regarding future failures. For the last few years, there has been an increasing interest on the decision making algorithms triggered by failure predictions, especially in production engineering. From the presented above state-of-the-art analysis we can make a similar conclusion as in [16], i.e.: machines without vibrations in the industrial environment are something non-existent. During machining operations, these vibrations are directly linked to the problems in systems having rotating or reciprocating parts, such as bearings, engines, gear boxes, shafts, turbines and motors. The vibration analysis has proved to be a measure for any cause of inaccuracy in manufacturing processes and components, or any maintenance decisions related to the machine. However, we should remember that a vibration signal analysis, on which most researchers are focused, is not the only one existing condition monitoring technique. Technological machines and machine tools as their typical example are equipped with several different sensors. The data gathered with them could be alternatively used in effective structural health monitoring. These aspects should be additionally discussed from the perspective of the internet of things-based intelligent decision support systems need [17], as a tool for data processing in manufacturing. With regard to the above presented research gaps and research challenges related to them, contributions of this work are twofold. An advanced milling machine multi-sensor system as a condition monitoring tool was presented. Its integration with the proposed signal processing method based on decision trees

with different types and levels of wavelets for the cutter reliability decision-making process was presented as well. The assumed indicators of achieving the research goal are: low calculation time and data processing complexity, a universal analytical tool, data gathered directly reflecting typical industrial production, and finally a high accuracy model to assess the condition of the cutter state in real-time.

The research presented in the article is based on the cooperation of the authors with the aviation industry, in which providing the quality of the manufactured components, including aircraft engine components, is a critical factor due to a potential threat to the lives of the aircraft passengers that is connected to this issue. In turn, the final and inter-process quality control introduces significant costs. What's more, it doesn't provide a full guarantee of the quality that, in many cases, could be only obtained by carrying out destructive tests. Therefore, it is purposeful to perform the works that will enable the development of the methods which increase the effectiveness of a technological process supervision, even at the expense of installing additional sensors, including the construction of machining handle instruments with the built-in sensors of e.g. force and torque. Component machining, including large components, always requires the use of appropriate instruments which position and hold the machining component to the machine tool workstation. The aviation industry is open to designing the instruments in a way to allow for installing in them appropriate sensors as soon as it is possible to achieve the benefits mentioned above.

The aim of the research covered in this article is to develop an effective and dedicated measurement system for monitoring critical machining procedures implemented in the aviation industry with the use of a single type tool. The proposed solution allows to achieve high efficiency for a particular machining procedure with the limited solution generality. The article consists of the introduction, followed by a chapter describing the experimental setup and data processing with the use of the selected techniques. Finally, prediction errors achieved with the use of different signal sources and data processing methods were presented and compared.

2. Experimental evaluation

In the first phase of the research presented, an advanced milling machine multi-sensor system was designed and constructed. It was assumed that the data had to be collected from different signal sources for the purpose of health monitoring and later on for the RUL prediction. The system should not only be universal from the research perspective but should also conform to the industrial conditions. For the wide analytical purposes a typical industrial milling machine working in real industrial conditions was equipped with such sensors as (Fig. 1): accelerometers collecting signals from the lower spindle bearing, upper spindle bearing, Z axis, upper motor bearing and lower motor bearing; an acoustic emission sensor, 3-axis force and torque sensor, spindle velocity and spindle load sensor.

The data collected with the use of this milling machine and multi-sensor condition monitoring system were used for the previously presented research results [24, 25]. Their aim was to apply vibration and acoustic signals analysis in health monitoring, cutter state classification or its remaining useful life prediction. For the purpose of the presented in the this article sensor system and signal processing integration for cutter reliability decision-making process, we have assumed that the data collected from the 3-axis force and torque sensor can be also applied. It may be used as an effective alternative to the typical vibration signal based health monitoring and the RUL prediction.

2.1. Experimental setup and data description

The main goal of the experiment was to collect the data describing the cutter state during a milling process. The state of the cutter was categorized into two classes: sharp and blunt. The experiment was carried out on an industrial Haas VM-3 CNC

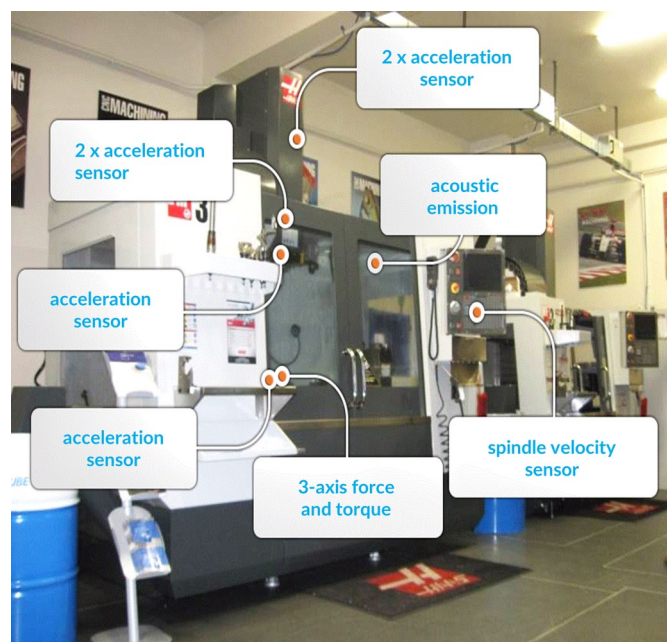


Fig. 1. Industrial CNC milling machine with a set of condition monitoring sensors [51]

machine. This machine is equipped with a 12,000 RPM direct drive spindle. The rotational speed of the spindle during machining was equal to 860 rpm. A multi-component CL16 ZEPWN sensor was used for the tests. The sensor enables force measurement in the range of 10 kN and torque measurement in the range of 1 kNm. The accuracy class of the sensor is 0.5, and the sensitivity is 1mV/V. The following signals were collected from the multi-component sensor: signals from the force sensor (P1x, P2y, P3z) and torque (M1x, M2y, M3z). A platform for rapid prototyping of intelligent diagnostic systems was used to collect data during milling experiments [51]. The platform includes Beckhoff industrial computer, an EtherCAT-based distributed I/O system. A hard disk of the engineering workstation was used to store the gathered data, collected in the real time with a sampling interval of 2 ms. The duration of the signal buffer stored in one file was 640 ms. During the experiments the data were collected from various real production tasks in the milling process on the machine.

2.2. Data processing

During the experiment a set of the collected data included 2172 observations. The data were gathered from the force sensor (signals: P1x, P2y, P3z) and torque (signals: M1x, M2y, M3z). These data were analysed in accordance with the methodology used to discover knowledge from the measurement database. The knowledge discovery in databases is a process of which the main task is a comprehensive data analysis, starting from the proper understanding of the problem under study, through the data preparation, execution and analysis of appropriate models, up to their evaluation. Then, the identified information is transformed into the knowledge that can be used to build decision support systems [3, 9]. In this paper the knowledge discovery process was divided into three stages: data pre-processing, data mining (processing), analysis of the results and evaluation of the created models (post-processing) (Fig. 2).

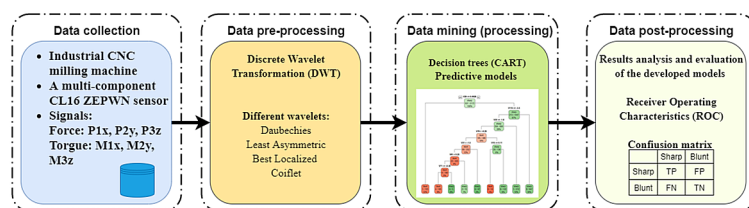


Fig. 2. Data processing methodology

In the first stage, the data obtained with the use of discrete wavelet transformation (DWT) were pre - processed. Different types and levels of wavelet were used [10]. The second stage of the knowledge discovery process was data mining. According to [26], which gives an overview of the methods used at this stage, primarily the methods related to the statistical data analysis, artificial intelligence and machine learning can be applied. Generally, these methods can be divided as follows:

- classic statistical methods, which include, among others: linear regression, multiple regression, analysis of variance,
- methods based on the use of artificial intelligence, machine learning and deep learning; for example: classification trees, regression trees, random forests, artificial neural networks, genetic algorithms, evolutionary algorithms, fuzzy sets, rough sets, enhanced and fuzzy trees, support vector machines and Bayes classifiers.

The authors of the aforementioned work [26] after analysing the results obtained in many publications, point out that in the case of large data sets, the methods from the second group are the most effective and most often used for data processing. In this paper, decision trees for data processing were used. In many publications, i.e. [7, 38], this method is widely used and it is considered as one of the best data mining algorithms. The results of their application in various research areas indicate their advantages such as: easy and transparent data interpretation, the ability to identify variables importance and ability to analyse large amounts of data [39, 41].

The third stage of the knowledge discovery was the interpretation and evaluation of the developed models. Receiver Operating Characteristics (ROC) was used as a tool to help to analyse the performance of predictive models. This method is often recommended for assessing the quality of models [8, 15, 36, 37].

2.2.1. Discrete Wavelet Transformation (DWT)

The data gathered from the force sensor (P1x, P2y, P3z) and torque (M1x, M2y, M3z) were preprocessed with the use of the wavelet analysis. The wavelet transformation is based on wavelet functions. Wavelet functions are irregular, asymmetric and, most of all, they are not periodic. The main goal of the wavelet transformation consists in the decomposition of the tested signal into component functions. Instead of harmonics, wavelet functions are used with a different scale (scale / frequency) and position (time / space) [10, 12]. The wavelet coefficients describe the extent to which the wavelet function is with a certain scale and position is similar to the considered signal fragment. The wavelet transformation consists in determining the coefficients for wavelets of various scales and positions.

Let \mathbb{N} denote a set of natural numbers, \mathbb{R} - set of real numbers, \mathbb{Z} - set of integer numbers. Let $\{x_t\}_{t \in \mathbb{Z}}$ be a time series and $\Psi(t)$ orthogonal wavelet basis - mother wavelet and $\phi(t)$ denotes the scaling function (father wavelet) corresponding to wavelet Ψ . For any $j \in \mathbb{Z}$ we define a sequences $\{\Psi_{jk}\}_{k \in \mathbb{Z}}$ and $\{\phi_{jk}\}_{k \in \mathbb{Z}}$ as follows:

$$\Psi_{jk}(t) = \frac{1}{2^{j-1}} \Psi\left(\frac{t}{2^j} - k\right) \quad (1)$$

and

$$\phi_{jk}(t) = \frac{1}{2^{j-1}} \phi\left(\frac{t}{2^j} - k\right) \quad (2)$$

Then the time series we can present as:

$$x_t = \sum_{k=-\infty}^{\infty} c_{jk} \phi_{jk}(t) + \sum_{i=-\infty}^j \sum_{k=-\infty}^{\infty} d_{ik} \Psi_{ik}(t) \quad (3)$$

where c_{jk} is a scaling coefficient, d_{ik} is a detailed coefficient.

In many cases we perform a wavelet transformation for a time series with a finite number of observations $\{x_t\}_{1 \leq t \leq n}$. A decomposing level j meets the condition $1 \leq j \leq m = \max\{s \in \mathbb{N} : 2^s \leq n\}$. To simplify, we assume that $n = 2^s$.

From the equations (1) and (2) we can see, that $\Psi_{jk}(t)$ and $\phi_{jk}(t)$ take non-zero values on the interval $[2^j k, 2^j(k+1)]$. From above the time series $\{x_t\}_{1 \leq t \leq n}$ we can present as follows:

$$x_t = \sum_{k=0}^{\frac{n}{2^j}-1} c_{jk} \phi_{jk}(t) + \sum_{i=0}^j \sum_{k=0}^{\frac{n}{2^i}-1} d_{ik} \Psi_{ik}(t) \quad (4)$$

for $1 \leq j \leq m$. Based on the equation (4) we see that the time series $\{x_t\}_{t \in \mathbb{Z}}$ can be presented in different forms due to the level $j \in \mathbb{Z}$.

According to [17], we define the time series projection operator $\{x_t\}_{1 \leq t \leq n}$ for the level j in the base $\{\phi_{jk}(t)\}_{0 \leq k \leq \frac{n}{2^j}-1}$:

$$P^j x_t = \sum_{k=0}^{\frac{n}{2^j}-1} c_{jk} \phi_{jk}(t) \quad (5)$$

More about the DWT can be found in [11, 35, 44].

2.2.2. Decision trees

Decision trees were used to develop predictive models for the processed signals. Decision trees are a family of data mining and machine learning methods that can be used for both classification and regression tasks. The classification task is performed for a variable characterized by a predetermined set of possible states or values, otherwise it is defined as a regression task. Decision trees use different algorithms. In this study, the CART (Classification and Regression Trees) algorithm was used, as presented in [5]. CART splits the observation sample for the target variable as a binary tree structure with non-intersecting subsamples called nodes, according to specific rules.

The construction criteria are used to stop the tree growth and to avoid the model overfitting. These include: a minimum number of observations in the parent node, a minimum number of observations in the child node, tree depth, a cross-validation type, reaching the specified error type and others. In the case of machine learning, the standard recommendation is to use 10 fold cross-validation. The resulting model includes all target cases classified in the terminal nodes of the tree. In order to classify a given data set with the help of decision trees, the conditions should be formulated in such a way as to obtain the greatest gain of information or the smallest Gini index. Therefore, the process of selecting an attribute is based either on the Gini index or on obtaining information [13, 19]. The Gini index is a measure used to measure the frequency with which the randomly selected items would be misclassified. The Gini index is defined as follows [18, 20]:

$$Q_G(m) = \sum_{j=1}^s p_{mi} (1 - p_{mi}) = 1 - \sum_{j=1}^s p_{mi}^2 \quad (6)$$

where p_{mi} is a conditional probability for j -th class in a node, s -a number of classes. In node m with n_m observations the conditional probability for j -th class is equal to:

$$p_{mi} = \frac{\#\{y = c_i : x \in R_m\}}{n_m} \quad (7)$$

2.2.3. Receiver Operating Characteristics (ROC) analysis

Receiver Operating Characteristics (ROC) indicators were used as a tool to help to determine the performance of predictive models. The ROC curve is a graph characteristic for a given classifier, showing TP (True Positives) and FP (False Positives) values on the Y and X axes. Classification errors are defined as FP (False Positives) and FN (False Negatives). They mean appropriately classifying objects from the positive to negative class and assigning cases from the negative to positive class. The values of TP, TN, FP and FN create the confusion matrix presented in Table 1.

Table 1. Confusion matrix

| Predicted classes | Real classes | |
|-------------------|---------------------|---------------------|
| | Positive | Negative |
| Positive | TP (True positive) | FP (False positive) |
| Negative | FN (False negative) | TN (True Negative) |

Based on the confusion matrix (Table 1), the following assessment indicators were used to assess the quality of classification models analysing the results from most of the classifiers of machine learning [14, 29, 36, 42, 46]:

- Accuracy (Acc), which is determined as the sum of TP and TN, it indicates that the results are correctly classified to all the analysed data. This indicator evaluates the prediction ability of the model:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

True Positive Rate (TPR) is the rate that determines the fraudulent free transactions classified as fraudulent:

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

- True Negative Rate (TNR) is the rate that determines the fraudulent free transactions classified as legitimate:

$$TNR = \frac{TN}{TN + FP} \quad (10)$$

- Positive Predictive value (PPV) is an indicator that describes the relationship between the number of true positives and the total number of positives: true positives and false positives:

$$PPV = \frac{TP}{TP + FP} \quad (11)$$

- Negative Predictive Value (NPV) is an indicator that describes the relationship between the number of true negatives and the total number of negatives: true negatives and false negatives:

$$NPV = \frac{TN}{TN + FN} \quad (12)$$

- Prevalence (PV) is an indicator that determines the frequency of occurrence of the distinguished class:

$$PV = \frac{TP + FN}{TP + TN + FP + FN} \quad (13)$$

- Detection Rate (DR) is an index that measures the ratio of true positives to the total number of predictions:

$$DR = \frac{TP}{TP + TN + FP + FN} \quad (14)$$

- Detection Prevalence (DPV) is an index defined as the number of predicted positive cases divided by the total number of predictions:

$$DPV = \frac{TP + FP}{TP + TN + FP + FN} \quad (15)$$

The ROC analysis is most often used to show how a change in the threshold value of a classifier affects its ability to classify. Using the ROC analysis, it is possible to select an optimal threshold value, also known as the cut-off point. Looking at the ROC curve only in this context, performing the ROC analysis would make sense only for a model that gives a numerical value on the output indicating the degree of belonging to the class (scoring). The ROC curve can also be used as a measure of the quality of a classifier by determining the Area under Curve (AUC) [29, 42].

3. Results and discussion

The main aim of the task was to recognize if the cutter was blunt or not, based on the observation of the signals obtained from sensors. From sensor for each signal P1x, P2y, P3z, M1x, M2y and M3z the sequence contained 320 observations was created. The sample realisation of signals are presented in Figure 3.

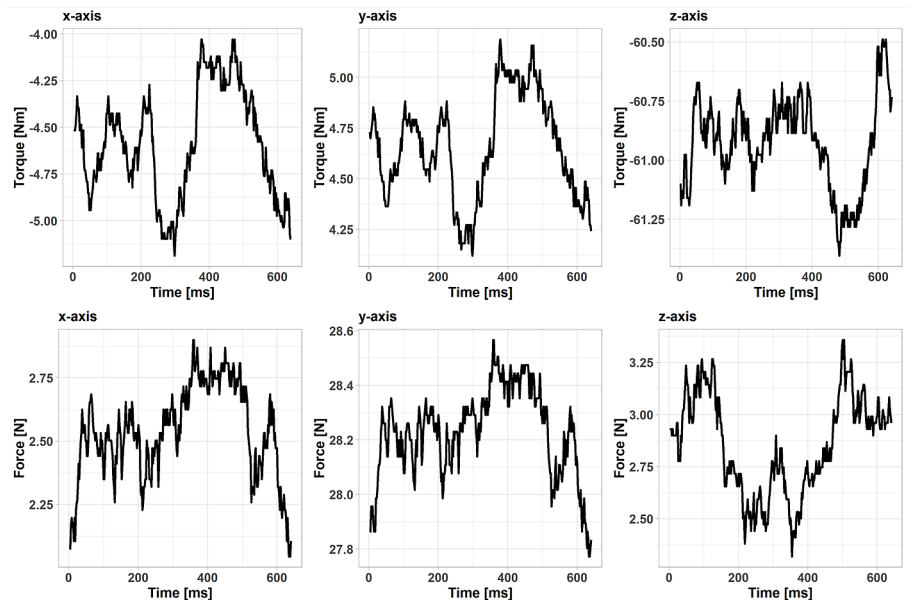


Fig. 3. Sample realisation of P1x, P2y, P3z, M1x, M2y and M3z signals

To analyse the relationship between the main characteristics of data and cutter state, the statistical analysis was performed. The Kruskal-Wallis test for hypothesis testing was used and the basic statistics were analysed. The Table 2 and Table 3 present the basic signal statistics for the cutter state.

Table 2. The basic signal statistics for a sharp cutter

| Signals | M1x | M2y | M3z | P1x | P2y | P3z |
|---------|------------|------------|------------|-----------|-----------|-----------|
| min | -0.1171427 | -0.0228960 | -0.6303399 | 0.0010815 | 0.2567350 | 0.0149973 |
| max | 0.0029251 | 0.1296336 | -0.5869904 | 0.0713316 | 0.3263204 | 0.0405637 |
| mean | -0.0529555 | 0.0403286 | -0.6181026 | 0.0248948 | 0.2874076 | 0.0282353 |
| std | 0.0310903 | 0.0302851 | 0.0062023 | 0.0172853 | 0.0095725 | 0.0055035 |
| 0.25% | -0.0803710 | 0.0146124 | -0.6218952 | 0.0100079 | 0.2840958 | 0.0235235 |
| 0.5% | -0.0514923 | 0.0402213 | -0.6191194 | 0.0209323 | 0.2882440 | 0.0283904 |
| 0.75% | -0.0248435 | 0.0671787 | -0.6148348 | 0.0389972 | 0.2938729 | 0.0329296 |

Table 3. The basic signal statistics for blunt cutter

| | M1x | M2y | M3z | P1x | P2y | P3z |
|-------|------------|------------|------------|-----------|-----------|-----------|
| min | -0.1439127 | -0.0656999 | -0.6350875 | 0.0030977 | 0.2707327 | 0.0058511 |
| max | 0.0510880 | 0.1436246 | -0.5717976 | 0.1140984 | 0.4079226 | 0.0521933 |
| mean | -0.0511779 | 0.0389308 | -0.6186331 | 0.0421332 | 0.3625294 | 0.0280621 |
| std | 0.0480753 | 0.0476963 | 0.0080547 | 0.0259049 | 0.0452849 | 0.0108821 |
| 0.25% | -0.0882373 | 0.0005646 | -0.6243047 | 0.0233395 | 0.3643719 | 0.0192413 |
| 0.5% | -0.0545385 | 0.0386228 | -0.6188666 | 0.0355519 | 0.3847081 | 0.0277085 |
| 0.75% | -0.0106254 | 0.0772895 | -0.6136879 | 0.0590801 | 0.3919459 | 0.0366029 |

Table 4. Kruskal-Wallis test results for torque and force signals

| | M1x | M2y | M3z | P1x | P2y | P3z |
|-------------|-----------|-----------|-----------|----------|----------|----------|
| chi-squared | 0.0619675 | 0.5523552 | 2.1087337 | 250.9589 | 687.9325 | 0.715012 |
| p-value | 0.8034128 | 0.4573570 | 0.1464605 | 0.0000 | 0.0000 | 0.397785 |

The Kruskal-Wallis test to compare the mean of the received torque and force signals was used. The test results are presented in the Table 4.

Analyzing the above data, it should be noted that at the significance level $\alpha = 0.01$ for M1x, M2y, M3z and P3z signals, there are no reason for rejecting the null hypothesis, that there is no statistically significant difference between the mean values for the sharp and blunt cutters (p-value > 0.01). That's why it should be noted that on the basis of the mean of the received signals it is not possible to determine the condition of the cutter. Therefore, signals were pre-processed by the application of a wavelet analysis.

For the possible wavelet a data set $D = \{(w_i, y_i)\}_{1 \leq i \leq n}$, was defined, where for i -th sample the value $y_i \in \{0, 1\}$ denotes the cutter state, but $w_i \in \mathbb{R}^m$ denotes the vector of predictors based on wavelet pre-processing. When the cutter was sharp then we put $y_i = 0$, otherwise if the cutter was blunt then $y_i = 1$. For designing a decision tree the data set which contains 2172 samples was used. For the chosen wavelet and filtering level $l \in \mathbb{N}$ the signals from sensors were pre-processed, i.e. the same preprocessing was applied to the observation sequences from P1x, P2y, P3z, M1x, M2y and M3z signals.

Thus, for each sample the sequences of approximation coefficients $\{c_{lk}^{j,s}\}_{1 \leq k \leq n}$ and detail coefficients $\{d_{lk}^{j,s}\}_{1 \leq k \leq n}$ were estimated, where $1 \leq j \leq 2172$ and $s \in \{P1x, P2y, P3z, M1x, M2y, M3z\}$.

For the signal decomposition the following different wavelets were applied:

- Daubechies 2,4,6,8,10,12,14,16,18,20;
- Least Asymmetric 8,10,12,14,16,18,20;
- Best Localized 14,18,20;
- Coiflet 6,12,18,24,30.

For the detail coefficients the mean $m_j^s = \sum_{k=1}^n d_{lk}^{j,s}$ and variance $S_j^s = \sum_{k=1}^n (d_{lk}^{j,s} - m_j^s)^2$ were determined. The vector of predictors was defined as:

$$w_j = \left\{ m_j^{P1x}, S_j^{P1x}, \{c_{lk}^{j,P1x}\}_{1 \leq k \leq n}, m_j^{P2y}, S_j^{P2y}, \{c_{lk}^{j,P2y}\}_{1 \leq k \leq n}, \dots, m_j^{M3z}, S_j^{M3z}, \{c_{lk}^{j,M3z}\}_{1 \leq k \leq n} \right\} \quad (16)$$

The set $D = \{(w_i, y_i)\}_{1 \leq i \leq 2172}$, where $y_i \in \{0, 1\}$, $w_j \in \mathbb{R}^{6(n+2)}$, is a learning set based on which classification trees were designed. The realizations of some features in the dataset for sharp and blunt cutters differ significantly. The density function and the distribution for one of the analyzed features are shown in Figure 4.

Significance of differences between sharp and blunt cutters based on possible feature which distribution is presented on figure 4 was confirmed by Kolmogorov-Smirnov and Kruskal-Wallis tests. For Kolmogorov-Smirnov test the statistic D is equal 0.7297983, for Kruskal-Wallis test the statistic χ^2 is equal 935.8446284. For both tests p - value is to approximately 0. Hence, at the significance level of 0.01, it should be assumed that the values of the presented feature for sharp and blunt cutters differ significantly.

The following tree construction criteria were used: a classification tree (method = 'class') and the following parameters that control the tree designing procedure: a complexity parameter (cp = 0.005), a minimum number of observations that have to exist on the node in order to be able to attempt a split (minsplit = 7), a number of variables competing at the output (maxcompete = 10), a number of surrogate variables (maxsurrogate = 10), a method of determining which surrogate variables will be used (usesurrogate = 2), and a maximum depth

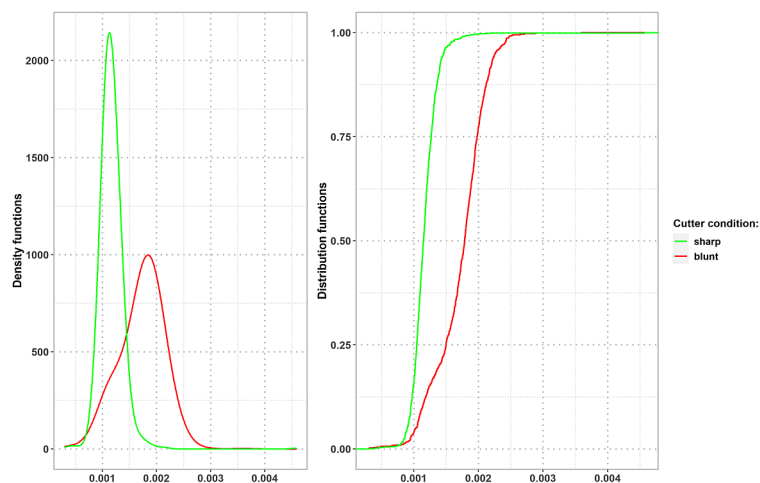


Fig. 4. Analysis of the distribution of an exemplary feature

of a tree (maxdepth = 7). A decision tree was generated for the defined parameters.

In order to evaluate the quality of recognition, the following ratios were estimated: accuracy, sensitivity (True Positives Rate), specificity (True Negatives Rate), positive predictive value, precision, negative predictive value, prevalence, detection rate and detection prevalence. Additionally, 10-fold cross validation was done. For each wavelet, the number of variables used in the training set (n.var) and the number

of variables used in the tree (n.used) were checked. In order to carry out the cross validation procedure, the learning set was divided into 10 portions. The classification tree was created every time based on the training set containing 9 portions. However, the accuracy was determined based on the test set containing only one portion. Each time the test set was different. For the obtained accuracy sequence the mean and standard deviation were estimated. The values of these parameters were attached to Table 5 as Acc.cv and Acc.sd respectively. The obtained results are presented in Table 5.

When analysing the results presented in Table 5, it should be noted that the first indicator (accuracy - Acc) shows that the highest value was obtained for the Daubechies 20 wavelets at level $l = 4$ (Acc = 0.9926). On the other hand, the lowest Acc value was obtained for the Coiflet 30 wavelet, level $l = 3$ (Acc = 0.9797). The results for the sensitivity (TPR) of the classifiers look similar. The ability to detect objects from the selected class is the highest for the Daubechies 20 wavelets at level $l = 4$ (TPR = 0.9904) and the lowest for Coiflet 30 wavelets level $l = 3$. The analysis of the TNR and PPV indicator shows its highest value for the Coiflet 6 wavelet at level $l = 5$. The highest probability of belonging of an object to the category recognized by the classifier as a not distinguished class in the actual non-displayed class (NVP) was obtained for Daubechies 20 wavelets at level $l = 4$. Though, the number of predicted positive cases (DPV) was obtained for Best Localized 14 wavelets at level $l = 4$. The value of the PV indicator for all the analysed wavelets was at a comparable level, that is ≈ 0.4314 . Fig-

Table 5. The values of prediction models quality indicators

| | level | n.var | n.used | Acc | TPR | TNR | PPV | NPV | PV | DR | DPV | Acc.cv | Acc.sd |
|---|-------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| d2 | 5 | 120 | 9 | 0.9862 | 0.9808 | 0.9903 | 0.9871 | 0.9855 | 0.4314 | 0.4231 | 0.4286 | 0.9630 | 0.0104 |
| d4 | 5 | 120 | 7 | 0.9834 | 0.9691 | 0.9943 | 0.9923 | 0.9769 | 0.4314 | 0.4180 | 0.4213 | 0.9713 | 0.0067 |
| d6 | 5 | 120 | 7 | 0.9843 | 0.9723 | 0.9935 | 0.9913 | 0.9792 | 0.4314 | 0.4194 | 0.4231 | 0.9722 | 0.0115 |
| d8 | 5 | 120 | 7 | 0.9802 | 0.9616 | 0.9943 | 0.9923 | 0.9715 | 0.4314 | 0.4148 | 0.4180 | 0.9706 | 0.0090 |
| d10 | 5 | 120 | 8 | 0.9853 | 0.9723 | 0.9951 | 0.9935 | 0.9793 | 0.4314 | 0.4194 | 0.4222 | 0.9669 | 0.0096 |
| d12 | 4 | 168 | 10 | 0.9890 | 0.9829 | 0.9935 | 0.9914 | 0.9871 | 0.4314 | 0.4240 | 0.4277 | 0.9623 | 0.0122 |
| d14 | 4 | 168 | 8 | 0.9848 | 0.9744 | 0.9927 | 0.9902 | 0.9808 | 0.4314 | 0.4203 | 0.4245 | 0.9685 | 0.0130 |
| d16 | 4 | 168 | 8 | 0.9866 | 0.9829 | 0.9895 | 0.9861 | 0.9871 | 0.4314 | 0.4240 | 0.4300 | 0.9719 | 0.0100 |
| d18 | 4 | 168 | 8 | 0.9834 | 0.9701 | 0.9935 | 0.9913 | 0.9777 | 0.4314 | 0.4185 | 0.4222 | 0.9663 | 0.0092 |
| d20 | 4 | 168 | 12 | 0.9926 | 0.9904 | 0.9943 | 0.9925 | 0.9927 | 0.4314 | 0.4273 | 0.4305 | 0.9684 | 0.0121 |
| la8 | 5 | 120 | 6 | 0.9853 | 0.9755 | 0.9927 | 0.9902 | 0.9816 | 0.4314 | 0.4208 | 0.4250 | 0.9768 | 0.0080 |
| la10 | 5 | 120 | 8 | 0.9894 | 0.9808 | 0.9960 | 0.9946 | 0.9856 | 0.4314 | 0.4231 | 0.4254 | 0.9795 | 0.0100 |
| la12 | 4 | 168 | 8 | 0.9862 | 0.9808 | 0.9903 | 0.9871 | 0.9855 | 0.4314 | 0.4231 | 0.4286 | 0.9735 | 0.0107 |
| la14 | 4 | 168 | 7 | 0.9848 | 0.9712 | 0.9951 | 0.9934 | 0.9785 | 0.4314 | 0.4190 | 0.4217 | 0.9708 | 0.0112 |
| la16 | 4 | 168 | 7 | 0.9843 | 0.9701 | 0.9951 | 0.9934 | 0.9777 | 0.4314 | 0.4185 | 0.4213 | 0.9677 | 0.0085 |
| la18 | 4 | 168 | 9 | 0.9876 | 0.9808 | 0.9927 | 0.9903 | 0.9855 | 0.4314 | 0.4231 | 0.4273 | 0.9689 | 0.0126 |
| la20 | 4 | 168 | 7 | 0.9816 | 0.9648 | 0.9943 | 0.9923 | 0.9738 | 0.4314 | 0.4162 | 0.4194 | 0.9705 | 0.0071 |
| bl14 | 4 | 168 | 7 | 0.9857 | 0.9840 | 0.9870 | 0.9829 | 0.9878 | 0.4314 | 0.4245 | 0.4319 | 0.9747 | 0.0116 |
| bl18 | 4 | 168 | 7 | 0.9820 | 0.9658 | 0.9943 | 0.9923 | 0.9746 | 0.4314 | 0.4167 | 0.4199 | 0.9677 | 0.0115 |
| bl20 | 4 | 168 | 10 | 0.9871 | 0.9808 | 0.9919 | 0.9892 | 0.9855 | 0.4314 | 0.4231 | 0.4277 | 0.9658 | 0.0101 |
| c6 | 5 | 120 | 7 | 0.9908 | 0.9840 | 0.9960 | 0.9946 | 0.9880 | 0.4314 | 0.4245 | 0.4268 | 0.9782 | 0.0096 |
| c12 | 4 | 168 | 7 | 0.9820 | 0.9658 | 0.9943 | 0.9923 | 0.9746 | 0.4314 | 0.4167 | 0.4199 | 0.9659 | 0.0101 |
| c18 | 4 | 168 | 9 | 0.9885 | 0.9829 | 0.9927 | 0.9903 | 0.9871 | 0.4314 | 0.4240 | 0.4282 | 0.9694 | 0.0088 |
| c24 | 3 | 276 | 8 | 0.9834 | 0.9691 | 0.9943 | 0.9923 | 0.9769 | 0.4314 | 0.4180 | 0.4213 | 0.9650 | 0.0107 |
| c30 | 3 | 276 | 7 | 0.9797 | 0.9594 | 0.9951 | 0.9934 | 0.9700 | 0.4314 | 0.4139 | 0.4167 | 0.9657 | 0.0076 |
| Legend: Prediction model with the highest Acc value Prediction model with the lowest Acc value | | | | | | | | | | | | | |

Figure 5 shows the decision tree for the wavelets with the highest value of the accuracy coefficient (Daubechies 20 wavelets for the level $l = 4$).

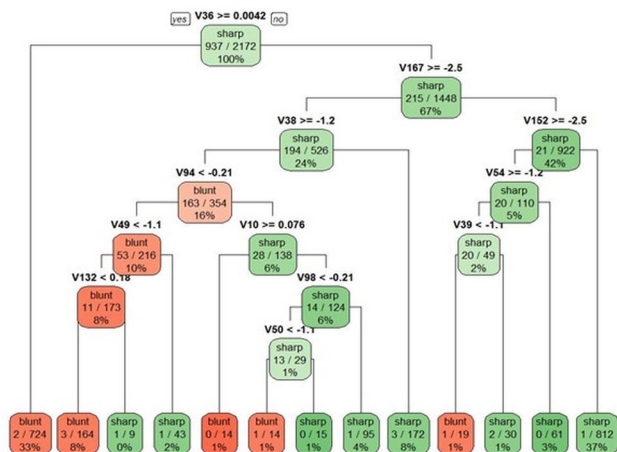


Fig. 5. A decision tree for the wavelets with the highest value of the accuracy coefficient (Daubechies 20 wavelets for the level $l = 4$)

Analysing Figure 5 it should be noted that with the defined 168 variables only 12 were used (Table 2) for the tree construction. The developed tree has 12 split nodes and 13 terminal nodes, and, thus, it generates 13 decision rules defining the cutter state. The ranking of the variables importance was used to build the tree in the training set for Daubechies 20 wavelets for the level $l = 4$ and presented in Figure 6. The highest values indicate the largest variable influence on the cutter state. In this case, from 168 variables used (Table 2) the most important variables are: V38, V54, V45, V47, V43, V46, V39, V40, V41, V36 and V92. The importance of the variable determines the participation of the variable in the created decision tree. Importantly, the specific meaning of a variable applies only to the analysed decision tree for which it was determined.

Variables included in the decision tree nodes do not necessarily mean its high importance. It can be observed that among these most important variables, only V38 and V39 were included in the analysed tree out of all 12 variables (nodes) in the tree. The variables which are to be included in the tree largely depend on the set of variables and its specificity. Input fields that contain relevant information may not be included in the decision tree and, thus, the quality of the forecast will not be affected. The analysis of the importance of the variables allowed to identify those input variables that have the greatest impact on creating the decision tree, and, thus, have an impact on the condition of the cutter state.

Table 6 presents the confusion matrix for the Daubechies 20 wavelets level $l = 4$. The sharp cutter is assumed to be a negative case (N), while a blunt cutter is a positive case (P). The confusion matrix analysis shows that 16 out of 2172 analysed variants were incorrectly classified, which means that the prediction error is $\approx 0.74\%$. This value indicates a very high predictive ability of the developed classifier.

On the other hand, the highest value of the accuracy indicator (Acc.cv = 0.9795) after the application of a 10-fold cross-validation was obtained for the Least Asymmetric 10 wavelets for the level $l = 5$. Figure 7 shows a decision tree created for the training set for the coefficients obtained on the basis of data processing using Least Asymmetric 10 wavelets.

Analysing Figure 7 it can be noted that with the defined 120 variables only 7 were used (Table 5) for the tree construction. The developed tree has 7 split nodes and 8 terminal nodes. It means that 8 decision rules define the cutter state. Table 7 shows the analysis results of the cutter state using a 10-fold cross-validation. In one of the analysed cases, the training set contained 1947 records, while the test set contained 225.

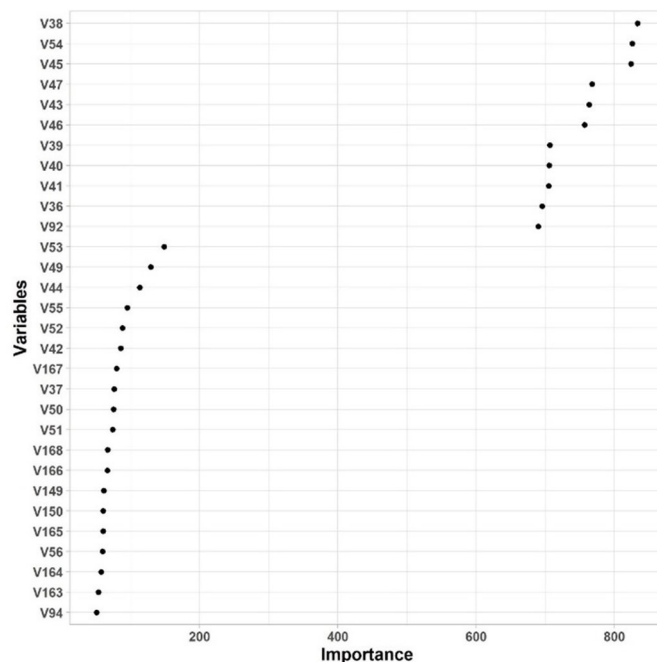


Fig. 6. The ranking of variable importance for a decision tree (Daubechies 20 wavelets for the level $l = 4$)

Table 6. Confusion matrix for classification tree designed on from wavelets Daubechies 20

| | Reference | | |
|------------|-----------|-------|-------|
| | State | Blunt | Sharp |
| Prediction | Blunt | 928 | 7 |
| | Sharp | 9 | 1228 |

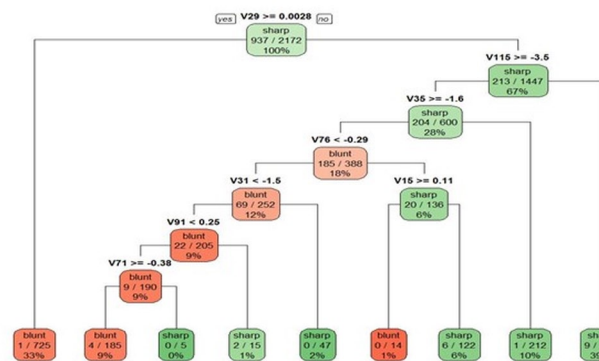


Fig. 7. A decision tree with the highest accuracy value after 10-fold cross-validation (Least Asymmetric 10 wavelets for level $l=5$)

Table 7. The chosen confusion matrix for the classification tree designed on the coefficients obtained from wavelets Least Asymmetric 10

| | Reference | | |
|------------|-----------|-------|-------|
| | State | Blunt | Sharp |
| Prediction | Blunt | 91 | 1 |
| | Sharp | 3 | 130 |

The analysis of the confusion matrix shows that 4 out of 225 set records analysed variants were incorrectly classified. Moreover, the lowest Acc.std value (Acc.std = 0.0067) (Table 2) was obtained for Daubechies 4 wavelets at level $l = 5$, which means that the changes in the Acc indicator value with the 10-fold cross-validation were the

smallest. It should be assumed that the predictive model for these wavelets is the most stable. However, the value of the accuracy indicator for this model was only $\text{Acc} = 0.9834$, which means the prediction error is $\approx 1.7\%$.

4. Conclusions

Technological machines designed for the Industry 4.0 applications, among which are also machine tools, are advanced mechatronic systems equipped with several sensors. The data gathered from them are usually used for diagnostic, monitoring and other purposes including their components remaining useful life prediction, condition classification, or both of them. Typical condition monitoring techniques (temperature, vibration or acoustic signal analysis) play an important role as data sources and indicators of a developmental failure, and having a wide range of different applications. Among them, a vibration signal analysis is usually applied as a measure for any cause of inaccuracy in manufacturing processes and components, or any maintenance decisions related to the machine. On the other hand, technological machines are equipped with several different sensors, from which the data gathered could be alternatively used in effective structural health monitoring. That is why, the aim of this article was to verify how effective appropriate data processing of such alternative signals collected from the multi-component sensor: signals from the force sensor and torques could be. All this with a strong relation to the expected solutions necessary in digital transformation, which will help to eliminate current barriers such as heterogeneous data streams that cannot be well processed to realize the automated decision support due to the lack of strong analytic capabilities. Another research challenge in this area should also be considered, that is the development of predictive data analytics techniques in order to aggregate and process the sensor data to assist in the maintenance operations or scheduling. It requires advanced information analytics for the networked machines that will finally be able to perform more efficiently and collaboratively. Vast research is conducted in this area. However, it is mainly theoretical considerations where new methods or mathematical models are usually verified only with the use of simulation data. Although there are many works investigating SHM or RUL in production engineering, they are usually limited to small and academic problems. Monitoring smart structures poses a big challenge in terms of fault or damage detection, due a huge amount of noisy data collected from many sensors on a periodic basis.

For the purpose of the presented sensor system and signal processing integration for a cutter reliability decision-making process, we assumed that the data collected from the 3-axis force and torque sensor can be also used as an alternative to a typical vibration signal based health monitoring and the RUL prediction, while integrated with machine learning techniques that are regarded as a powerful solution. An industrial milling machine multi-sensor system as a condition monitoring tool was presented. Its integration with the proposed signal processing method based on decision trees with different types and levels of wavelets for the cutter reliability decision-making process was a part of the research results discussed. In the first stage, the data gathered were pre-processed with the use of discrete wavelet transformation. The main goal of the wavelet transformation consists in the decomposition of the tested signal into component functions. Different types and levels of wavelet were used. Next, decision trees (a family of data mining and machine learning methods that can be used for both classification and regression tasks) were applied for data processing in order to develop predictive models for the processed signals. The third stage of the knowledge discovery was the interpretation and evaluation of the developed models. Receiver Operating Characteristics (ROC) was used as a tool to assess the performance of the developed predictive models. The presented confusion matrix for the classification tree designed on the coefficients obtained from wavelets Least Asymmetric 10 allowed to achieve a prediction error equal to 1.7%. On the other hand, much better results were achieved in

the case of the classification tree designed with the use of Daubechies 20 wavelets. Only 16 out of 2172 analysed variants were incorrectly classified, which means that the prediction error was equal to 0.74%.

The data gathered during the same industrial production process but coming from other sensors were also analysed for classification and prediction on earlier research stages. The comparison of the results achieved before with these, presented in the current work, will allow to verify the research hypothesis i.e. is it possible to use the data collected from the 3-axis force and torque sensor as an alternative to the typical vibration signal based health monitoring and RUL prediction? In [25] the prediction was evaluated by the SVM application. Cutter condition identification was done by registering and processing vibroacoustic data, in conjunction with torque measurement using a three-axis sensor mounted in the chuck. A correlation analysis, which is related to the spectral analysis, was used to identify the parametric property of an vibroacoustic signal, but torque signals were identified as ARIMA models. This information was used to create a data set. Additionally, for the prediction based on SVM the modified kernel function as a linear combination of kernels representing the acoustic signal and torque data was used. The prediction error achieved was equal to 2.1 %. In [9] SVM was applied only for the preprocessed vibroacoustic signals. In this case the achieved prediction error was equal to 2.6%. In [24] the prediction was assessed by a logistic regression application into the preprocessed vibroacoustic signals. The classification error was obtained at the level of 8.6%. The comparison of the results achieved previously and in the current analysis indicates a very high predictive ability of the analysed tree and alternative condition monitoring data source. A novel approach for a predicting tool remaining useful life was also proposed by Li et al. [27], who emphasize that most current approaches for the predicting tool RUL are based on historical failure and truncation data, while for the new types of tools or when a similar tool has just been launched, such failure and truncation data are limited or even unavailable. In order to address this problem, a novel method for the prediction of the tool RUL using limited data was proposed and, for this purpose, a time window was constructed to track the tool condition using sensor data, with its size to be dynamically adjusted according to the wear factor and increase rate. Then, a deep bidirectional long short-term memory neural network in which sequential data are predicted and smoothed by forwards and backwards directions respectively, was developed to encode temporal information and identify long-term dependencies. On this basis, multi-step ahead rolling predictions were employed to predict the tool RUL. The presented results [27] show that with this method it is possible to predict the tool RUL. However, its weakness stems from the time consuming and complicated multi-step framework of the proposed prediction algorithm. In addition, this algorithm is also quite sensitive to changes in tool working conditions. The mean absolute error and root mean square error of the method proposed by [27] were 0.1130 and 0.1592. They are much higher than prediction errors achieved in this study.

To sum up, the novelty aspect and most important achievements of the research results presented in the article are:

1. It was proved with the use of real world industrial production process data that the 3-axis force and torque sensors can be considered as a data source alternative to the typical vibration signal for health monitoring and RUL prediction, while integrated with adequate pre- and post-processing methods.
2. The possible application of different types and levels of wavelets for signal processing with their efficiency analysis in industrial condition monitoring were presented and discussed.
3. Different predictive models were developed with the use of decision trees for the signals processed with various types and levels of wavelets proving their ability to accurately trace a tool condition.
4. The ROC analysis was used to identify the most stable predictive model and the model with the lowest prediction error selection method.

5. The prediction error of the proposed method is lower than those for the previously proposed approaches evaluated, while the time necessary to achieve high accuracy predictions and analysis complexity as well as the associated cost are much lower. Limited calculation time and data processing complexity reduction are significant results of the proposed method.

In Industry 4.0 vision, data digitalization and data processing are expected to bring major changes in manufacturing in general, and the spread of novel technologies will enable a stepwise increase of productivity in manufacturing companies. From this perspective, the described milling machine multi-sensor system and the proposed data

processing model may be considered as a decision-making process tool in determining cutting tools service life, extending the time of their effective use in a production process, making this way the replacement time as optimal as possible.

Acknowledgement:

The research was partially financed in the framework of the project Lublin University of Technology - Regional Excellence Initiative, funded by the Polish Ministry of Science and Higher Education (contract no. 030/RID/2018/19)".

References

1. Ahamd A, Paul A, Din S, Rathore MM, Choi GS, Jeon G. - Multilevel data processing using parallel algorithms for analyzing Big Data in high-performance computing. *International Journal of Parallel Programming* 2018; 46: 508-527, <https://doi.org/10.1007/s10766-017-0498-x>.
2. Arrazola P, Özel T, Umbrello D, Davies M, Jawahir I. - Recent advances in modelling of metal machining processes. *CIRP Annals* 2013; 62: 695-718, <https://doi.org/10.1016/j.cirp.2013.05.006>.
3. Borucka A, Wiśniowski P, Mazurkiewicz D, Świderski A. - Laboratory measurements of vehicle exhaust emissions in conditions reproducing real traffic. *Measurement* 2021; 174: 108998, <https://doi.org/10.1016/j.measurement.2021.108998>.
4. Bousdekis A, Lepenioti K, Apostolou D, et al. - Decision making in predictive maintenance: Literature review and research agenda for Industry 4.0. *IFAC-PapersOnLine* 2019; 52: 607-612, <https://doi.org/10.1016/j.ifacol.2019.11.226>.
5. Breiman L, Friedman JH, Olshen RA, Stone CJ. - *Classification and Regression Trees*. Chapman and Hall/CRC: Boca Raton, FL, USA 1984.
6. Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L, Zdeborova L. - Machine learning and the physical sciences. *Reviews of Modern Physics* 2019; 91: 045002.
7. Choi S, Battulga L, Nasridinov A, Yoo K-H. - A Decision Tree Approach for Identifying Defective Products in the Manufacturing Process. *International Journal of Contents* 2017; 13: 57-65, <https://doi.org/10.5392/IJoC.2017.13.2.057>.
8. Costa EP, Lorena AC, Carvalho ACPLF, Freitas AA. - A review of performance evaluation measures for hierarchical classifiers. *Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop*, AAAI Press 2007: 82-196.
9. Dargan S, Kumar M, Ayyagari MR, Kumar G. - A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of Computational Methods in Engineering* 2020; 27: 1071-1092, <https://doi.org/10.1007/s11831-019-09344-w>.
10. Daubechies I. - Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics* 1998; 41: 909-996.
11. Daubechies I. - *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics 1992, <https://doi.org/10.1137/1.9781611970104>.
12. Edwards T. - *Discrete Wavelets Transform: Theory and Implementation*. Stanford University 1991.
13. Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R. - Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications* 2014; 41: 1937-1946, <https://doi.org/10.1016/j.eswa.2013.08.089>.
14. Fawcett T. - An introduction to ROC analysis. *Pattern Recognition Letters* 2006; 27: 861-874, <https://doi.org/10.1016/j.patrec.2005.10.010>.
15. Fawcett T. - Using rule sets to maximize ROC performance. *Proceedings 2001 IEEE International Conference on Data Mining ICDM 2001: 131-138*, <https://doi.org/10.1109/ICDM.2001.989510>.
16. Goyal D, Pabla B. - The vibration monitoring methods and signal processing techniques for structural monitoring: a review. *Archives of Computational Methods in Engineering* 2016; 23: 585-594, <https://doi.org/10.1007/s11831-015-9145-0>.
17. Guo Y, Wang N, Xu ZY, Wu K. - The internet of things-based decision support system for information processing in intelligent manufacturing using data mining technology. *Mechanical Systems and Signal Processing* 2020; 142: 106630, <https://doi.org/10.1016/j.ymssp.2020.106630>.
18. Hastie T, Tibshirani R, Friedman J. - *The elements of statistical learning*. Springer-Verlag New York Inc 2009, <https://doi.org/10.1007/978-0-387-84858-7>.
19. Hssina B, Merbouha A, Ezzikouri H, Erritali M. - A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications* 2014; 4: 13-19.
20. James G, Witten D, Hastie T, Tibshirani R. - *An introduction to statistical learning*. Springer-Verlag GmbH 2013, <https://doi.org/10.1007/978-1-4614-7138-7>.
21. Jasiulewicz-Kaczmarek M, Antosz K, Wyczółkowski R, Mazurkiewicz D, Sun B, Qian C, Ren Y. - Application of MICMAC, Fuzzy AHP and Fuzzy TOPSIS for Evaluation of the Maintenance Factors Affecting Sustainable Manufacturing. *Energies* 2021; 14(1436): 1-31, <https://doi.org/10.3390/en14051436>.
22. Jiang S, Sun SY. - Stability analysis for a milling system considering multi-point-contact cross-axis mode coupling and cutter run-out effects. *Mechanical Systems and Signal Processing* 2020; 141: 106452, <https://doi.org/10.1016/j.ymssp.2019.106452>.
23. Koch W. - *Tracking and sensor data fusion. Methodological framework and selected applications*. Springer Verlag, Berlin 2014.
24. Kozłowski E, Mazurkiewicz D, Żabiński T, Prucnal S, Sęp J. - Assessment model of cutting tool condition for real-time supervision system. *Eksploracja i Niezawodność – Maintenance and Reliability* 2019; 21: 679-685, <https://doi.org/10.17531/ein.2019.4.18>.
25. Kozłowski E, Mazurkiewicz D, Żabiński T, Prucnal S, Sęp J. - Machining sensor data management for operation-level predictive model. *Expert Systems with Applications* 2020; 159: 1-22, <https://doi.org/10.1016/j.eswa.2020.113600>.
26. Lepenioti K, Bousdekis A, Apostolou D, Mentzas G. - Prescriptive analytics: literature review and research challenges. *International Journal of Information Management* 2020; 50: 57-70, <https://doi.org/10.1016/j.ijinfomgt.2019.04.003>.
27. Li H, Wang W, Li Z, Dong L, Li Q. - A novel approach for predicting tool remaining useful life using limited data. *Mechanical Systems and Signal Processing* 2020; 143: 1086832, <https://doi.org/10.1016/j.ymssp.2020.1086832>.

28. Liu R, Kothuru A, Zhang S. - Calibration-based tool condition monitoring for repetitive machining operations. *Journal of Manufacturing Systems* 2020; 54: 285-293, <https://doi.org/10.1016/j.jmsy.2020.01.005>.
29. Matthews BW. - Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 1975; 405: 442-451, [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9).
30. Mazurkiewicz D. - Empirical and analytical models of cutting process of rocks. *Journal of Mining Science* 2000; 36: 481-486, <https://doi.org/10.1023/A:1016620810143>.
31. Nath C. - Integrated tool condition monitoring systems and their applications: a comprehensive review. *Procedia Manufacturing* 2020; 48: 852-863, <https://doi.org/10.1016/j.promfg.2020.05.123>.
32. Neugebauer R, Denkena B, Wegener K. - Mechatronic systems for machine tools. *CIRP Annals* 2007; 56: 657-686, <https://doi.org/10.1016/j.cirp.2007.10.007>.
33. Nouri M, Fussell BK, Ziniti BL, Linder E. - Real-time tool wear monitoring in milling using a cutting condition independent method. *International Journal of Machine Tools and Manufacture* 2015; 89: 1-13, <https://doi.org/10.1016/j.ijmachtools.2014.10.011>.
34. Pelayo GU, Trejo DO. - Model-based phase shift optimization of serrated end mills: minimizing forces and surface location error. *Mechanical Systems and Signal Processing* 2020; 144: 106860, <https://doi.org/10.1016/j.ymssp.2020.106860>.
35. Percival DB, Walden AT. - Wavelet methods for time series analysis. Cambridge University Press 2000.
36. Powers DM. - Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technology* 2011; 2: 37-63.
37. Provost F, Fawcett T, Kohavi R. - The case against accuracy estimation for comparing classifiers, *Proceedings of the ICML-98*, Morgan Kaufmann, San Francisco 1998: 445-453.
38. Raghavan V. - Application of decision trees for integrated circuit yield improvement. In the 13th Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference, Advancing the Science and Technology of Semiconductor Manufacturing ASMC 2002; 02CH37259: 262-265, <https://doi.org/10.1109/ASMC.2002.1001615>.
39. Ronowicz J, Thommes M, Kleinebudde P, Krysiński J. - A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm. *European Journal of Pharmaceutical Sciences* 2015; 73: 44-48, <https://doi.org/10.1016/j.ejps.2015.03.013>.
40. Schuld M, Sinayski I, Petruccione F. - An introduction to quantum machine learning. *Contemporary Physics* 2015; 56: 172-185, <https://doi.org/10.1080/00107514.2014.964942>.
41. Shao Q, Rowe RC, York P. - Comparison of neurofuzzy logic and decision trees in discovering knowledge from experimental data of an immediate release tablet formulation. *European Journal of Pharmaceutical Sciences* 2017; 31: 129-136, <https://doi.org/10.1016/j.ejps.2007.03.003>.
42. Sokolova M, Japkowicz N, Szpakowicz S. - Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: Sattar A., Kang B. (eds) *AI 2006: Advances in Artificial Intelligence*. AI 2006, Lecture Notes in Computer Science 2006; 4304, https://doi.org/10.1007/11941439_114.
43. Vamsi I, Sabareesh GR, Penumakala PK. - Comparison of condition monitoring techniques in assessing fault severity for a wind turbine gearbox under non-stationary loading. *Mechanical Systems and Signal Processing* 2019; 124: 1-20, <https://doi.org/10.1016/j.ymssp.2019.01.038>.
44. Walnut DF. - An introduction to wavelet analysis. Springer Nature 2004.
45. Wang Y, Zheng L, Wang Y. - Event-driven tool condition monitoring methodology considering tool life prediction based on industrial internet. *Journal of Manufacturing Systems* 2021; 58: 205-222, <https://doi.org/10.1016/j.jmsy.2020.11.019>.
46. Wu X, Kumar V, Ross J, Quinlan J, et al. - Top 10 algorithms in data mining. *Knowledge and Information Systems* 2008; 14: 1-37, <https://doi.org/10.1007/s10115-007-0114-2>.
47. Xi S, Cao H, Chen X. - Dynamic modelling of spindle bearing system and vibration response investigations. *Mechanical Systems and Signal Processing* 2018; 114: 486-511, <https://doi.org/10.1016/j.ymssp.2018.05.028>.
48. Zhang C, Qian Y, Dui H, Wang S, Chen R, Tomovic MM. - Opportunistic maintenance strategy of a Heave Compensation System for expected performance degradation. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2021; 23(3): 512-521, <http://doi.org/10.17531/ein.2021.3.12>.
49. Zhang C, Yao X, Zhang J, Jin H. - Tool condition monitoring and remaining useful life prognostic based on wireless sensor in dry milling operations. *Sensors* 2016; 16: 795, <https://doi.org/10.3390/s16060795>.
50. Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. - Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing* 2019; 115: 213-237, <https://doi.org/10.1016/j.ymssp.2018.05.050>.
51. Żabiński T, Mączka T, Kluska J. - Industrial Platform for Rapid Prototyping of Intelligent Diagnostic Systems. *Trends in Advanced Intelligent Control, Optimization and Automation - Polish Control Conference*, eds. W. Mitkowski, J. Kacprzyk, K. Oprządkiewicz, P. Skruch 2017: 712-721, https://doi.org/10.1007/978-3-319-60699-6_69.