Springer Series in Reliability Engineering

Yu Liu
Dong Wang
Jinhua Mi
He Li *Editors*

Advances in Reliability and Maintainability Methods and Engineering Applications

Essays in Honor of Professor Hong-Zhong Huang on his 60th Birthday

Springer

Springer Series in Reliability Engineering

Series Editor

Hoang Pham, Industrial and Systems Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

Today's modern systems have become increasingly complex to design and build, while the demand for reliability and cost effective development continues. Reliability is one of the most important attributes in all these systems, including aerospace applications, real-time control, medical applications, defense systems, human decision-making, and home-security products. Growing international competition has increased the need for all designers, managers, practitioners, scientists and engineers to ensure a level of reliability of their product before release at the lowest cost. The interest in reliability has been growing in recent years and this trend will continue during the next decade and beyond.

The Springer Series in Reliability Engineering publishes books, monographs and edited volumes in important areas of current theoretical research development in reliability and in areas that attempt to bridge the gap between theory and application in areas of interest to practitioners in industry, laboratories, business, and government.

Now with 100 volumes!

Indexed in Scopus and EI Compendex

Interested authors should contact the series editor, Hoang Pham, Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854, USA. Email: **hopham@soe.rutgers.edu, or Anthony Doyle, Executive Editor, Springer, London. Email: anthony.doyle@springer.com.**

Yu Liu · Dong Wang · Jinhua Mi · He Li
Editors

Advances in Reliability and Maintainability Methods and Engineering Applications

Essays in Honor of Professor Hong-Zhong Huang on his 60th Birthday

Springer

Editors
Yu Liu
Center for System Reliability and Safety
University of Electronic Science
and Technology of China
Chengdu, Sichuan, China

Jinhua Mi
Center for System Reliability and Safety
University of Electronic Science
and Technology of China
Chengdu, Sichuan, China

Dong Wang
Department of Industrial Engineering
and Management, School of Mechanical
Engineering
Shanghai Jiao Tong University
Shanghai, China

He Li
Centre for Marine Technology and Ocean
Engineering (CENTEC), Instituto Superior
Técnico
Universidade de Lisboa
Lisbon, Portugal

ISSN 1614-7839 ISSN 2196-999X (electronic)
Springer Series in Reliability Engineering
ISBN 978-3-031-28858-6 ISBN 978-3-031-28859-3 (eBook)
https://doi.org/10.1007/978-3-031-28859-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

It is my pleasure to write the foreword for this well-edited book: *Advances in Reliability and Maintainability Methods and Engineering Applications*. Contributed chapters in the book cover the latest trends in the areas of reliability, maintainability, and safety engineering as well as applications in various fields including aviation, power systems, wireless networks, and mechanical engineering systems. The chapters have succeeded in achieving a fine balance between theory and practice.

I am indeed delighted to know that this book was brought on a special occasion of the 60th birthday of a dear friend, Prof. Hong-Zhong Huang, who has contributed immensely to the field of reliability engineering. He truly deserves this special honor and tribute by some of his former Ph.D. students and collaborators.

Advances in failure rate and degradation modeling of mechanical systems, reliability of wireless body networks and aircraft systems, inspection policies of complex systems, opportunistic maintenance of wind turbines, and proactive health management are some of the highlights of the book. I am sure that students, research scholars, scientists, engineers, and practitioners of reliability engineering will greatly benefit from this book.

I strongly recommend this book for its comprehensive coverage on the advances of reliability and maintainability engineering and their practical applications.

Many best wishes to you, Prof. Hong-Zhong Huang, on your 60th birthday!

December 2022

Dr. Hoang Pham
Editor-in-Chief
Springer Book Series in Reliability Engineering,
Fellow of the IEEE, AAIA, IISE
Distinguished Professor
Rutgers University
Piscataway, NJ, USA

Preface

This book aims at collecting state-of-the-art research works from globally renowned scholars in the field of reliability and maintainability to celebrate Prof. Hong-Zhong Huang's 60th birthday in June 2023. Professor Hong-Zhong Huang is Full Professor at the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China, where he also serves as Director of the Center for System Reliability and Safety. He received his B.Sc. degree at Wuhan University in 1983, M.Sc. degree at Chongqing University in 1988, and Ph.D. degree at Shanghai Jiao Tong University in 1999. He has held visiting appointments with several universities in the USA, Canada, and Asia. He has authored or coauthored more than 320 journal papers and eight books in the fields of reliability engineering, optimization design, prognostics, and health management. He is ISEAM Fellow, Technical Committee Member of ESRA, Co-editor-in-Chief for the *International Journal of Reliability and Applications*, and Editorial Board Member of several international journals. In the past 28 years as Graduate Advisor, he supervised and co-supervised 18 postdoctoral research fellows, 68 Ph.D. students, and 182 M.Sc. students.

In recognition of Prof. Huang's excellent and global contributions to the reliability and maintainability community, this book is composed of 24 invited chapters from 12 countries including China, the USA, the UK, Canada, Germany, Portugal, Poland, Singapore, Japan, South Korea, Chile, and the Kingdom of Saudi Arabia and 43 institutions. All these chapters were contributed by worldwide renowned scholars and research groups. The layout of this book is as follows:

In Chapter "Multi-criteria Based Selection of Ship-Based Ballast Water Treatment Technologies", a multi-criteria-based selection of ship-based ballast water treatment technologies is presented. Chapter "A Two-Phase Sampling Approach for Reliability-Based Optimization in Structural Engineering" reports a two-phase sampling approach to address reliability-based optimization problems in structural engineering. In Chapter "Moment Estimation-Based Method of Motion Accuracy Reliability Analysis for Industrial Robots", a novel computational framework is proposed to comprehensively evaluate the reliability for kinematic positioning and

trajectory accuracy of industrial robots. In Chapter "Reliability of Wireless Body Area Networks", the reliability model of wireless body area networks is introduced. In Chapter "Sensitivity Estimation of Markov Reward Models and Its Applications to Component Importance Analysis", Markov chain-based component-wise sensitivity analysis is put forth to evaluate component importance measures without any system structure function. Chapter "Failure Rate Modeling of Mechanical Components and Systems" introduces failure rate modeling of mechanical components and systems. Chapter "Statistical Reliability Modeling and Analysis for Repairable Systems with Bathtub-Shaped Intensity Function" reports statistical reliability modeling and analysis for repairable systems with bathtub-shaped intensity function. In Chapter "Multi-state Signatures for Multi-state Systems with Binary/Multi-state Components", signatures for multi-state systems with binary/multi-state components are derived. Chapter "Comprehensive Reliability of Aircraft Actuation System" presents the essential reliability characteristics of redundant aircraft actuation systems and creates a reliability evaluation method for non-similar redundancy actuation systems. Chapter "Integration of Reliability Design, Installed Base, and After-Sales Services for System Availability" discusses a holistic approach to sustain system availability in an integrated product-service framework. Chapter "Use of Artificial Neural Networks to Enhance Container Port Safety Analysis under Uncertainty" proposes a modified failure mode effect analysis approach by using artificial neural networks to predict the operational risks of container terminals. Chapter "Usage of Failure Time and Repair Time for Optimization of Maintenance and Warranty Policy and Lemon Law Application" reports the usage of failure time and repair time for optimization of maintenance and warranty policy and Lemon Law application. Chapter "Reliability and Opportunistic Maintenance of Floating Offshore Wind Turbines" reviews the state-of-the-art methods and procedures for reliability and maintainability analysis of floating offshore wind turbines. Chapter "A Summary of Inspection Policies of One Shot Systems" summarizes inspection optimization models for one-shot systems. Chapter "Analysis for Influence of Maintenance and Manufacturing Quality on Reliability of Repairable Systems" reports analysis of the influence of maintenance and manufacturing quality on the reliability of repairable systems. Chapter "Quantification of Uncertainty of Warranty Claims" reviews warranty, introduces its different types, discusses possible causes of warranty claims, and provides an introductory overview of the approaches to modeling warranty claims. Manufacturing paradigm-oriented prognostics and health management methodologies for cyber-physical systems are introduced in Chapter "Manufacturing Paradigm-Oriented PHM Methodologies for Cyber-Physical Systems". In Chapter "Degradation Modeling and Residual Life Prediction Based on Nonlinear Wiener Process", degradation models based on nonlinear Wiener processes under univariate and multivariate situations are presented. In Chapter "System Reliability Models with Dependent Degradation Processes", various system reliability models with dependent degradation processes are introduced. In Chapter "A Study of Health State Transitions for Proactive Health Management", health state transitions for proactive health management are studied. In Chapter "Kalman Filter-Based Systems Approach for Prognostics and Health Management of Electric Motors", a Kalman

filter-based systems approach is presented for prognostics and health management of electric motors. Chapter "Exploratory Fault Detection with Multivariate Data: A Case Study on Engine Bearing" studies detection of impending bearing failures using in-situ field data. Chapter "Novel Approach to Prognostics and Health Management to Combine Reliability and Process Optimisation" provides a novel approach for prognostics and health management to combine reliability and process optimization. Chapter "Current Status and Prospects of Reliability Systems Engineering in China" presents a systematic overview of the evolution of reliability systems engineering in China and emphatically introduces its latest developments.

We would like to sincerely acknowledge all the contributors and friends who make this book possible. Our special thanks are given to Prof. Dong Ho Park at Hallym University, Prof. Carlos Guedes Soares at University of Lisbon Higher Technical Institute, Prof. Joe Mathew at Asset Institute, Prof. Krzysztof Kołowrocki at Gdynia Maritime University, Prof. Enrico Zio at Politecnico di Milano, Prof. Narayanaswamy Balakrishnan at McMaster University, Prof. David W. Coit at Rutgers University, Prof. Michael Beer at Leibniz Universität Hannover, Prof. Min Xie at City University of Hong Kong, Prof. Zili Wang at Beihang University, Prof. Jin Wang at Liverpool John Moores University, Prof. Tadashi Dohi at Hiroshima University, Prof. Won Young Yun at Pusan National University, Prof. Joo-Ho Choi at Korea Aerospace University, Prof. Loon Ching Tang at National University of Singapore, Prof. Sukjoo Bae at Hanyang University, Prof. Xu Han at Hebei University of Technology, Prof. Lifeng Xi and Prof. Ershun Pan at Shanghai Jiao Tong University, Prof. Lirong Cui at Qingdao University, Prof. Shaoping Wang at Beihang University, Prof. Bo Guo at National University of Defense Technology, Prof. Liyang Xie at Northeastern University, Prof. Renyan Jiang at Changsha University of Science and Technology, Prof. Liudong Xing at University of Massachusetts Dartmouth, Prof. Dariusz Mazurkiewicz at Lublin University of Technology, Prof. Tongdan Jin at Texas State University, Prof. Zaili Yang at Liverpool John Moores University, Prof. Shaomin Wu at University of Kent for their warm supports on this book, and all the coauthors. Last but not least, we would like to thank Prof. Hoang Pham at Rutgers University, who gave a warm and great support to the publication of this book. It is also indeed our pleasure working with Miss Sharmila Anbu and the Springer editorial team.

The four editors of this book have a close relationship with Prof. Huang, and they are Prof. Huang's former graduate students. On behalf of all his students, we would like to take the opportunity to wish Prof. Huang continuing success, happiness, and health in all coming years.

Chengdu, P.R. China — Dr. Yu Liu
Shanghai, P.R. China — Dr. Dong Wang
Chengdu, P.R. China — Dr. Jinhua Mi
Lisbon, Portugal — Dr. He Li
February 2023

Biography of Prof. Hong-Zhong Huang

Dr. Hong-Zhong Huang is Professor at the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China (UESTC). He serves as Director of the Center for System Reliability and Safety at UESTC. Professor Huang received his Ph.D. degree from Shanghai Jiao Tong University, China, in 1999. His research interests are reliability engineering, optimization design, prognostics, and health management. While reaching 60 years old in June 2023, Prof. Huang has achieved remarkable achievements. He has published more than 320 papers in journals, including *Reliability Engineering & System Safety*, *IEEE Transactions on Reliability*, *IISE Transactions*, *Annals of Operations Research*, and *Journal of Mechanical Design-Transactions* of the ASME. He has also published eight books and 10 book chapters and edited 18 international conference proceedings. About 68 of his patents have been granted. Professor Huang is Internationally Renowned Scholar, ISEAM Fellow, and Technical Committee Member of European Safety and Reliability Association (ESRA). He serves as Co-editor-in-Chief for the *International Journal of Reliability and Applications* and Editorial Board Member of several international journals (e.g., *Reliability Engineering & System Safety*, *International Journal of Metrology* and *Quality Engineering*).

Professor Huang graduated with distinction from the Wuhan University of Hydraulic and Electric Engineering (1979–1983), one of the Project 211 and a national key university, and merged with Wuhan University in 2000. He received his master's degree from Chongqing University as a top student in 1988 with the honor of "Excellent Postgraduate Student", under the supervision of Prof. Wenji Chen. As the part of a National "Seventh Five-Year Plan" Major Scientific and Technological Research Project, his master's thesis contributed to solving problems of strength calculation and weight optimization of large excavator booms. He won a special prize for the National Major Technical Equipment Achievements issued by the State Council of China. Supervised by Prof. Zongwu Hu (1996–1999), his Ph.D. dissertation, entitled "Study on Fuzzy Reliability Theory and Its Application to Mechanical Components and Systems", was awarded as one of the top 100 National Outstanding Doctoral Dissertations in China. As Postdoctoral Research Fellow (1999–2001), he

worked with Prof. Runfang Li in the State Key Laboratory of Mechanical Transmission at Chongqing University and was awarded an "Excellent Postdoctoral Fellow" by the People's Government of Sichuan Province.

He was employed by Southwest Electric Power Design Institute in Chengdu (1983–1985). As Institute's Technical Representative, he oversaw the construction of the Chongqing Huayingshan Power Plant. Due to his outstanding performance, he undertook the responsibility of designing a coal handling system for Yunnan Xiaolongtan Power Plant and an automatic coal unloading line for the dumper of Chongqing Baima Power Plant.

He worked for Southwest Jiaotong University and was appointed as Director of the Reliability and Intelligent Design Research Office (1988–1999). He received an early promotion to Associate Professor in 1994 and Professor in 1998. During this period, Prof. Huang was involved in the preliminary design of a rolling vibration test bench for whole locomotives and vehicles of the State Key Laboratory of Traction Power (the first largest in Asia and the second largest in the world). His research projects were supported by leading funding bodies in his subjects such as National Natural Science Foundation of China, China Postdoctoral Science Foundation, Science and Technology Development Fund of the Ministry of Railways, Sichuan Province Cross Century Distinguished Young Discipline Leader Fund, and Sichuan Province's First Batch of Academic and Technical Leader Training Fund.

His achievements focused on:

(1) Fuzzy methods in reliability engineering. He made critical comments on conventional reliability theories and discussed fuzzy reliability theories' generation, development, and application prospects (see Huang 1994).
(2) Applications of computational intelligence in mechanical engineering. He introduced the concept of fuzzy mechanical science and technology and established its theoretical framework (see Huang 1996).
(3) Design, finite element analysis, and optimization of products. It included CHT12 container cranes, railway rescue cranes, crane trolleys, ring crushers, impeller unloaders, and SH6600 light passenger cars. It also included the development of a tower crane safety evaluation expert system and a railway crane conceptual design expert system.

His paper "Research on fuzzy fault tree theories and applications" was awarded the Best Paper at the Asia Pacific Symposium on Occupational Safety and Health in 1993. In addition, many of his papers were granted the Excellent Paper Awards from the Sichuan Association for Science and Technology and Sichuan Mechanical Engineering Society.

He published three books in Beijing, *Reliability Theory and Application of Mechanical Transmission* (by China Science and Technology Press) in 1995, *The Principle and Application of Fuzzy Optimization in Mechanical Design* (by Science Press) in 1997, and *Fuzzy Design* (by China Machine Press) in 1999. He also published about 100 papers in such as *Microelectronics Reliability*, *Journal of Mechanical Engineering*, and *China Safety Science Journal*.

He was elected as Distinguished Young Scholar of Sichuan Province, the first batch of candidates for academic and technical leaders of Sichuan Province. His deeds were reported by Sichuan Daily. He was awarded the "Hongyu Excellent Teacher Award of Southwest Jiaotong University" in 1994 and the "Third Outstanding Young Teacher Award of Chengdu" in 1996. On July 26, 1995, he was honored by meeting with Presidents Zemin Jiang and Jintao Hu at the Great Hall of the People in Beijing. His teaching program entitled "A Series of Reforms of Mechanical Design Courses in Strengthening Practical Links and Improving Design Ability" was awarded the Second Prize of Excellent Teaching Achievements of Southwest Jiaotong University in the academic year 1991–1992. Moreover, he received the First Prize of Science and Technology Progress Award of the Sichuan Provincial Education Commission in 1997 for his project titled "Study on Reliability of Mechanical Transmission".

He served as Member of the editorial board of Modern Design Series, the 4th editorial board of *Journal of Machine Design* (August 1999), the Second Committee of the Reliability Engineering Branch of China Mechanical Engineering Society (October 1995), and Director of the Reliability Society of China Modern Design Method Research Society (December 1995).

He worked for the School of Mechanical Engineering at Dalian University of Technology (2000–2004), as Director of the Institute of Advanced Design Technology and Intelligent Control and Member of the Council of the University. Dr. Huang was Adjunct Professor at several universities including Wuhan University. In addition, he was Senior Visiting Scholar at the University of Alberta in Canada (2003–2004) and Visiting Professor at the National University of Singapore (November–December 2004). He was elected to the Excellent Young Teachers Funding Program of the Chinese Ministry of Education. He also served as Vice Chairman of the Mechanical Design Theory and Method Committee of the Mechanical Design Branch of the Chinese Mechanical Engineering Society. He was Member of several professional societies, including the Third Committee of the Reliability Engineering Branch of the Chinese Mechanical Engineering Society, the Fifth Committee of the Mechanical Design Branch of Chinese Mechanical Engineering Society, and the Teaching Committee of Mechanical and Electrical Subjects in Colleges and Universities of China Education Association of Machinery Industry. As Member of the fourth editorial board of *Journal of Machine Design*, he also served as General Chair of the National Modern Design Theory and Method Committee Council and Academic Meeting (2001), the National Reliability Academic Meeting and the Third Committee Council of Reliability Engineering Branch (2002), the Session Chair of the 5th International Conference on Frontiers of Design and Manufacturing (2002), and the Third Cross-Strait Workshop on Manufacturing Technology (2003).

Professor Huang founded the collaborative laboratory "Joint Research Office of Advanced Design and Control Technology" and "Joint Research Center for Advanced Technology of Construction Machinery" with the Taiyuan Heavy Machinery Co., Ltd., and Guangxi Liugong Group Co., Ltd., respectively, and served as Executive Deputy Director.

During this period, His research focused on fuzzy reliability and intelligent optimization. He received research grants from the National Natural Science Foundation of China, the National Natural Science Foundation of China Projects of International (Regional) Cooperation and Exchanges Program, the National Excellent Doctoral Dissertation Foundation of the Chinese Ministry of Education, Excellent Young Teachers Foundation Program of the Chinese Ministry of Education, Subprojects of the National Tenth Five-Year Science and Technology Research Project "Development of Automobile Safety Auxiliary Devices", and some Open Projects Fund by the State Key Laboratory. Cooperating with large backbone enterprises such as Guangxi Liugong Group Co., Ltd., Taiyuan Heavy Machinery Group Co., Ltd., Wafangdian Bearing Group Co., Ltd., and Liaohe Oilfield, he was committed to addressing reliability issues of products such as loaders, excavators, bearings, and petroleum machinery.

He also participated in writing the book *Chinese Mechanical Design Canon* and edited three conference proceedings. One of his papers, entitled "Fuzzy multiobjective optimization methods based on functional-link network", was awarded the Best Paper by the China Mechanical Engineering Society (2001), and his project "Design Theory of Mechanical Fuzzy Reliability" won the Second Prize of the Natural Science Award of Chinese Ministry of Education (2002).

Professor Huang has been at the University of Electronic Science and Technology of China (UESTC) since January 2005, as Dean of the School of Mechatronics Engineering (2006–2013). Currently, he is Director of the International Cooperation and Education Committee, Deputy Head of the Science and Technology Committee, Director of the Center for System Reliability and Safety, and Member of the University's Academic Committee.

He was Senior Visiting Scholar at Northwestern University in USA (2005–2006), Visiting Professor in the Center for PHM at the City University of Hong Kong (April–May 2010), as well as Visiting Professor at the University of Alberta (July–August 2010). He was appointed as Technical Advisor of XCMG Research Institute Co., Ltd. (July 2021–December 2022), and Consultant of Taiyuan Heavy Machinery Group Co., Ltd. (August 2022–July 2024). Due to his prominent achievements, he was elected as "Grassland Talent" in Inner Mongolia Autonomous Region, China (2016–2021), and Distinguished Professor at the Inner Mongolia University of Technology, China (2017–2020), where he was nominated as Honorary Dean of the School of Mechanical Engineering (2020–2023).

Professor Huang was also honored with a special government allowance expert by China's State Council and the National Science Fund for Distinguished Young Scholars (Type B). He takes leadership in many committees. He has been awarded as Model Worker, Academic Leader, Outstanding Expert with Extraordinary Contributions, and Leading Innovation Talent of the Sichuan Province.

Professor Huang served as Member of academic committees in several laboratories, including the State Key Laboratory of Mining Equipment and Intelligent Manufacturing, the Key Laboratory of Space Launch Site Reliability Technology, and the Key Laboratory of Equipment Support Engineering. He also served as Chairman of

academic committees at Hebei Provincial Key Laboratory of PHM, Inner Mongolia Autonomous Region Key Laboratory of Advanced Manufacturing Technology, etc.

He has been Fellow of the International Society of Engineering Asset Management (ISEAM Fellow) since 2010 and Member of the System Reliability Technical Committee of the European Safety and Reliability Society (ESRA) since 2007. He also served as Vice President of the Reliability Branch of the Operations Research Society of China (April 2013–December 2021), the Reliability Engineering Branch of the Chinese Mechanical Engineering Society (2018–present), the Reliability Branch of the Chinese Institute of Electronics (2019–present), and the Maintenance Professional Committee of Chinese Ordnance Society (2010–present). Meanwhile, he has been President of the Reliability Engineering Branch of the Sichuan Mechanical Engineering Society (2016–present), Vice Chairman of the PHM Professional Committee of Sichuan Mechanics Society (2019–present), and the Electronic Product Reliability and Quality Management Committee of Sichuan Institute of Electronics (2019–present).

Professor Huang has been appointed as Co-editor-in-Chief of the *International Journal of Reliability and Applications* (2016–present) and Associate Editor of the *International Journal of Metrology* and *Quality Engineering* (2022–present). Meanwhile, he serves in the editorial board of several journals, including *Reliability Engineering and System Safety* (2012–present), *Eksploatacja I Niezawodnosc-Maintenance and Reliability* (2012–present), *Applied Sciences* (2021–present), *Chinese Journal of Mechanical Engineering* (English, 2013–present), *Journal of Systems Engineering and Electronics* (2018–present), *International Journal of Reliability, Quality* and *Safety Engineering* (2007–2021). He also serves as Editorial Board Member of multiple Chinese journals and books.

Professor Huang was Founder or Co-founder of several international conference series on reliability, including the International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), the International Conference on Materials and Reliability (ICMR), and the International Conference on Maintenance Engineering (ICME), where he has served as General Chair, Co-chair, and Program Committee Chair. In addition, he has been invited to give 28 keynote speeches in both international and national conferences.

He has been Principal Investigator (PI) or participator in a number of government-sponsored projects like the National Natural Science Foundation Project of China, the "973 Program", the "863 Program", the National Science and Technology Major Project, and the National Key R&D Programs of China. In addition, Prof. Huang has held over dozens of company-sponsored projects such as Taiyuan Heavy Machinery Group Co., Ltd., Yuchai Group, SAIC-GM-Wuling Automobile, and Qiqihar No. 2 Machine Tool Group Co., Ltd.

He was awarded seven prizes for his remarkable contribution, including the Natural Science Award of the Ministry of Education and the Science and Technology Progress Award of Sichuan Province. He was granted the "William A. J. Golomski Award", an international academic award in the field of Reliability and Maintainability in California, USA, in January 2006. Many of his papers received

the Best Paper Awards both at international and national conferences. Since 2014, he has been recognized as Highly Cited Chinese Researcher (Elsevier) every year.

In the realm of education, Prof. Huang is also highly accomplished, as Member of several teaching steering committees, including the Mechanical Engineering Majors in Colleges and Universities of the Ministry of Education (2018–2022), the Industrial Engineering Majors in Colleges and Universities of Chinese Ministry of Education (2013–2017), the Mechanical Design Manufacturing and Automation Discipline of China Machinery Industry Education Association (2007–present), and the Sichuan Provincial Ordinary Undergraduate Colleges and Universities (Mechanical Engineering Major) (Vice President, 2021–2025). Apart from that, he served as Member of the Postgraduate Teaching Resources Construction Committee of Mechanical Engineering Discipline of Higher Education Press (2008–present). Professor Huang is Member of the Organizing Committee of the Sichuan Division of the National College Student Mechanical Innovation Design Competition (2006–2013) and the Textbook Editorial Committee of the Intelligent Manufacturing Series (2021–present). Also, he has been leading the UESTC's new engineering education project "Intelligent Equipment Reliability Summit Plan" (2022–2026), the Construction and Practice of the Mechanical and Electrical Innovation and Entrepreneurship Talent Training System of the Key Project of Sichuan Higher Education Talent Training Quality and Teaching Reform Project "Four Chains Integration" (2021–2023), and UESTC's postgraduate innovation training program "Innovative Talent Ability Improvement Plan for High-end Equipment Intelligence and High-Reliability Demand" (2021–2022). Two of his Ph.D. students received the HIWIN Doctoral Dissertation Award, the most prestigious recognition for Ph.D. students in the field of mechanical engineering. Given his distinguished accomplishments in education and teaching reform, he received the Second Prize in National Teaching Achievement and the First Prize in Sichuan Provincial Teaching Achievement Award (twice).

References

1. Huang, H. Z. (1994). A critical review of conventional reliability theories - Generation, development, and application prospects of fuzzy reliability theory. Journal of Machine Design, 11(3), 1–5, 47.
2. Huang, H. Z. (1996). Foundation of fuzzy mechanical science and technology - An important development direction of mechanical science in the 21st century. Journal of Mechanical Engineering, 32(3), 1–8.

Professor Hong-Zhong Huang at his Master's defense (April 1988)

Professor Hong-Zhong Huang (the first person from the right) with his Ph.D. advisor Prof. Zongwu Hu (the first person from the left) and the committee members of his doctoral defense (1998)

Professor Hong-Zhong Huang at the new product appraisal meeting of Guangxi LiuGong Machinery Co., Ltd. (March 28th, 2003)

Professor Hong-Zhong Huang with Prof. Wei Chen (Member of National Academy of Engineering) when he was a senior visiting professor at Northwestern University, USA (November 27th, 2004)

Professor Hong-Zhong Huang at ICME2006 (October 17th, 2006)

Professor Hong-Zhong Huang (the first person from the left in the front) visiting the University of Electro-Communications (Japan) with the former president of University of Electronic Science and Technology of China (April 4th, 2007)

Professor Hong-Zhong Huang with Ji Zhou (the former Minister of Education of China) at Tsinghua University (April 28th, 2012)

Professor Hong-Zhong Huang with some delegates of QR2MSE2014 (July 23th, 2014)

Professor Hong-Zhong Huang with his students (June 18th, 2017)

Professor Hong-Zhong Huang visiting Taiyuan Heavy Machinery Group Co., Ltd. (October 15th, 2018)

Professor Hong-Zhong Huang giving a welcome speech at ICMR2019, Jeju Island, South Korea. (November 28th, 2019)

Professor Hong-Zhong Huang giving a plenary speech at CSAA/IET AUS 2022, Nanchang, China. (August 19th, 2022)

Congratulatory Remarks

Professor Dong Ho Park
Honorary President, Korean Reliability Society
Korea

It is a great privilege for me to address congratulatory remarks for Prof. Hong-Zhong Huang who will celebrate his 60th birthday in June 2023. I would like to extend my sincere congratulations to him for keeping his health excellent, along with his outstanding achievements throughout his academic career in education and research. As I recall, I first met Prof. Huang at International Conference on Maintenance Engineering (ICME) held in Chengdu, China, in October 2006, and later we met again at the RAMS held in Las Vegas, USA, in 2008. Since then, we have been in close contacts as colleagues and friends for more than 15 years so far. On the occasion of my retirement in 2012, he even came to Korea to address a warm and unforgettable congratulatory remarks for me, which I am genuinely grateful for his hearty support and kindness. His love and energy for academic excellence and his gentle personality impressed me greatly and made me to have a deep respect for him.

Professor Huang has made a great contribution on several academic fields, especially on the field of maintenance and reliability engineering throughout his academic

career as a researcher and an educator. Numerous academic awards from several academic societies are the convincing evident of his outstanding academic achievements, and he surely is deserved for such a recognition for his life-long contribution. Professor Huang has been such an insightful and impactful leader, superb scholar, and warm and caring colleague, plus valued friend to many of us. He has been on so many fronts internationally and in his native land by giving utmost efforts to develop the area of maintenance and reliability engineering into a major academic field in the science. For international cooperation and mutual communication among worldwide researchers, he assumed a leading role to organize several international conferences, such as International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE) and International Conference on Materials and Reliability (ICMR), which have grown into remarkably productive international forums for many researchers to exchange their innovative ideas among them. More significant contribution he has made, in my opinion, is his invaluable efforts to produce a number of young and talented researchers by running Center for System Reliability and Safety Engineering at University of Electronic Science and Technology of China. Without his academic aspiration and consistent efforts to educate young students, such achievements of producing many excellent researchers would not have been possible.

During the course of many academic activities, we have had in many years, I have been very fortunate to get acquainted with a highly respected colleague and friend like Prof. Huang. Again, I would like to congratulate Prof. Huang for his 60th birthday after an exceptionally successful academic career with a good health. Although the age of 60 is no longer a young age, it does not necessarily mean the downhill of an active academic career, but rather the beginning of a new era for life.

I wish Prof. Huang, his family, and all his former students the most enjoyable and happy life for many more years to come. Most respectfully to all.

Congratulatory Remarks

Professor Enrico Zio
Centre de recherche sur les Risques et les Crises
Sophia Antipolis, France
Politecnico di Milano
Milan, Italy

It is more than 10 years since I first met Prof. Hong-Zhong Huang in beautiful Chengdu, China, at the 8th International Conference on Reliability, Maintainability and Safety (ICRMS2009) in 2009. Since then, I have had the opportunity to share many initiatives with Prof. Huang and the big and productive group that he has been able to establish in China to perform quality research in reliability engineering. Among these initiatives, I treasure also the opportunity to co-advise a visiting Ph.D. student from Prof. Huang's research group during the years 2016–2018, with excellent research outcomes and great personal satisfaction.

Through my professional relationship and personal friendship, I have learned to appreciate very much Prof. Hong-Zhong Huang as a colleague and as a friend. He is a worldwide renowned scientist in the field of reliability engineering, and the research

group led by him has contributed several relevant works related to system reliability modeling, maintenance optimization, and reliability-based design optimization.

Professor Huang's research group has devoted substantial efforts to developing new tools and approaches for tackling important reliability challenges in real-world engineering assets, such as machining tools, aeroengines, military vehicles, and industrial robots. As a result of their work, many influencing papers have been published in top journals, like *IEEE Transactions on Reliability*, *IISE Transactions*, *ASME Journal of Mechanical Design*, and others. The contributions of Prof. Huang and his research group to both academia and industry have been awarded international and national awards, like the William A. J. Golomski Award, ISEAM Fellow, Science and Technology Awards from the Chinese government, and others.

Professor Hong-Zhong Huang also made a great contribution to education and knowledge sharing for the entire reliability community of scientists, practitioners, and students, by founding the International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR^{2}MSE), the largest and most influencing event in Asia on these topics. In 2011, I was greatly honored to be invited to give a keynote lecture in the first event of this conference series in Xi'an. Already back then, the conference had attracted more than 200 participants from all over the world. To date, the annual event has been successfully held for 12 years and has continuously grown in participation, reaching a total number of participants of over 500 per year.

Finally, I like to conclude this (too) short congratulating note by:

- Thanking Prof. Hong-Zhong Huang for his friendship, of which I am honored.
- Thanking Prof. Hong-Zhong Huang for all opportunities of collaboration that he has offered me, for which I am truly grateful.
- Complimenting Prof. Hong-Zhong Huang for all that he has done in his professional life.
- Sincerely wishing my friend Hong-Zhong a more-than-nice (60th) birthday and a happy continuation of life full of personal joy and professional satisfaction.

Congratulatory Remarks

Professor Joseph Mathew
Chief Executive Officer, Asset Institute
Australia

It is my privilege to offer these congratulatory remarks to you Prof. Hong-Zhong Huang on reaching your 60th birthday milestone.

I offer my warmest congratulations on all your achievements, Prof. Huang and wish you even more success in the future.

I recall that I have known you since 2011 through the QR2MSE series of conferences which you initiated and grew as one of your many accomplishments. I have had the privilege of joining you and esteemed colleagues from around the world on several occasions in various venues in China and have enjoyable and very memorable experiences. No doubt your reputation in the field has attracted high-quality participation in the conference and which has resulted in numerous quality publications following these events. Of particular note was the opportunity to collaborate with you on organizing a joint QR2MSE conference with the 11th World Congress on Engineering Asset Management (WCEAM) in the spectacular venue of Jiuzhaigou,

China, in 2016. That conference has been etched in my mind as one of the most memorable of all the WCEAMs I have had the privilege of organizing.

No doubt, the accolades of your achievements are numerous and will come from many quarters around the world. One that I had personal involvement was your nomination to become a fellow of the International Society of Engineering Asset Management (ISEAM) in 2010—a nomination that is a result of highly notable and professional regard and standing in the field of reliability theory and application.

I would like to wish you a very happy 60th birthday Prof. Huang and I look forward to meeting you again in the near future.

Congratulatory Remarks

Professor Krzysztof Kołowrocki
President, Polish Safety and Reliability Association
Gdynia, Poland

Dear Sir
Prof. Hong-Zhong Huang
University of Electronic Science and Technology of China, Chengdu
The Founder of QR2MSE

Dear Professor
Distinguished Jubilee

On behalf of myself and the Board of the Polish Safety and Reliability Association, I would like to express my appreciation to the Professor for many years of scientific and didactic activity in the field of reliability. Your committed research, publication, and training activities as well as the development of new methods and tools have significantly contributed to raising the level of knowledge about reliability and safety in many companies and institutions and to the proper shaping of reliability

and safety culture in industry. This has brought measurable benefits in the development of young researchers, in ensuring the safety and health protection of industrial workers, economic development, and environmental protection.

We wish you a lot of health and satisfaction in your personal and professional life and further research achievements, in conducting the widely recognized development of young scientific staff and further managerial continuation of the internationally recognized international conference QR2MSE.

On the occasion of the 60th anniversary of birth.

Congratulatory Remarks

Professor David W. Coit
Rutgers University
Piscataway, NJ, USA

Prof. Hong-Zhong Huang has well served the engineering and reliability community very effectively in multiple ways with his great technical and research abilities and his effective leadership skills. This book and the associated chapters provide a meaningful tribute to his many achievements and his exemplary character. His abilities and achievements as a researcher alone would be sufficient for such a tribute, but his leadership and persistence in promoting our research interests in reliability and maintainability is even more important. As a researcher, he has few peers. His contributions and recognitions are so many. He has been a leader is conducting advanced research in condition monitoring, system reliability models, remaining life assessment, failure analysis, and other topics. His research papers always represent innovative new ideas that his students and other researchers can learn from and build upon. In addition to his achievements as a researcher, he has also been an academic leader and effective administrator in his positions of responsibility at the School of

Mechanical and Electrical Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, as well as visiting positions in USA, Canada, and other Asian countries. For many of us in the international research community, Prof. Huang is best known as the founder and leader of the incredibly successful QR2MSE conference held annually in different cities in China. I was fortunate to attend the first two offerings of this conference in Xi'an and Chengdu and attended several other times as well. The success of this conference from the very beginning is an amazing accomplishment and so impressive. Professor Huang has always been able to bring the very best researchers from China and all other the world to share their ideas and then to collaborate with attending students and professors. As a professor from Rutgers University in the USA, I had the opportunity to travel to China and learn so many new topics and be exposed to so many new ideas, as well as to share my own ideas and theories. Finally, I would like to thank Prof. Huang for being a good friend. The hospitality he personally provided to me and other visiting academics is very appreciated.

Congratulatory Remarks

Professor Min Xie
City University of Hong Kong
Hong Kong, China

I am pleased to have known Prof. Hong-Zhong Huang for over 20 years, and I fondly remember our first meeting when I was with National University of Singapore. Professor Huang visited me and we had interesting discussion on a number of topics of mutual interest. Since then, we have met many times at conferences and other occasions.

Professor Hong-Zhong Huang has dedicated himself to reliability research for over 35 years and developed many new reliability algorithms and tools aiming at addressing critical reliability problems in real-world engineering scenarios. These research outcomes have been worldwide recognized and received several international and national awards, such as William A. J. Golomski Awards and the Science and Technology Progress Award of National Defense.

Professor Hong-Zhong Huang has also dedicated himself to reliability education. He is the founder of the Center for System Reliability and Safety at the University

of Electronic Science and Technology of China (UESTC), China, and supervised 18 postdoctoral research fellows, 68 Ph.D. students, and 182 graduate students. The alumni are now serving very important roles in both academia and industry all over the world. I also have had opportunities to work directly with some of his bright and talented students.

Professor Hong-Zhong Huang's contributions to the reliability community have been worldwide recognized. He is the founder and general chair of the largest reliability events in Asian region, titled International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering. The events have been successfully held 12 times since 2011. He serves his editorial roles as the co-editor-in-chief of *International Journal of Reliability and Applications* and the editorial board member of several prestigious reliability journals such as *Reliability Engineering* and *System Safety*. He is also a fellow of International Society of Engineering Asset Management (ISEAM), the Technical Committee member of European Safety and Reliability Association (ESRA), and the vice president of four Reliability Societies of China.

I was impressed by Prof. Hong-Zhong Huang's accomplishments and personality, and this edited volume served as a nice gift on the occasion of his 60th birthday. I also take this opportunity to wish Prof. Hong-Zhong Huang a healthy life full of the achievements and joy.

Contents

Reliability Modelling, Analysis, and Design Optimization

Multi-criteria Based Selection of Ship-Based Ballast Water Treatment Technologies

Eugene Pam, Alan Wall, Zaili Yang, Eddie Blanco-Davis, and Jin Wang

Abstract The reality of selecting an acceptable ballast water treatment technology is a daunting task for end-users, due to availability of numerous treatment options and their efficacy in given ship-types and ballast voyages. Six treatment systems have been selected from the two generic treatment technology groups (physical solid liquid separation and disinfection), and are considered as the decision-making alternatives in the proposed model. The proposed model involves the application of the Technique for Order Performance by Similarity to the Ideal Solution (TOPSIS), in the decision-making analysis. The TOPSIS technique has been applied to obtain the performance ratings of the decision alternatives using linguistic terms parameterised with triangular fuzzy numbers. A sensitivity study is also conducted to identify the effects of changes in input data, and test the suitability of the developed model in decision-making analysis of ballast water treatment systems.

Keywords Multiple criteria · Ballast water · TOPSIS · Decision making

1 Introduction

Regulations D2 and D4 of the IMO International Convention for the Control and Management of Ships' Ballast Water and Sediments Ballast Water (2004), stipulate that all ships under construction in or after 2009 and having a ballast capacity between 1500 and 5000 m^3, must have ballast water treatment systems fitted to and used onboard with effect from January 1, 2009 [1]. Compliance to such IMO Regulations has propelled the development of numerous ballast water treatment technologies. The selection of a particular treatment system for a designated vessel or voyage route will have to be pre-determined by technical (safety of crew, ship and cargo), cost (production and running) and environmental (sustainability of the marine ecosystems) variables.

E. Pam · A. Wall · Z. Yang · E. Blanco-Davis · J. Wang (✉)
Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool L3 3AF, UK
e-mail: j.wang@ljmu.ac.uk

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_1

Evaluating these variables may not be straightforward due to inherent uncertainties and inadequacy of historical data. The choice of an appropriate ballast water treatment system can therefore, be a daunting task for both ship-owners and managers. Port states and/or regional regulatory authorities are also subject to decision-making problems, as they are expected to strike a balance between the sustenance of a pollution-free maritime environment, and the promotion of maritime trade of their countries/regions.

A novel model is developed in this paper to deal with multiple criteria decision-making (MCDM) problems associated with the analysis and selection of ballast water treatment systems, under a subjective group decision framework. A group decision-making problem arises when there are two or more individuals who, characterized by their perceptions, attitudes, motivations, and personalities, recognize the existence of a common problem and attempt to reach a collective decision [2]. In the developed model, fuzzy sets theory (FST), Analytic Hierarchy Process (AHP) and the Technique for Order Performance by Similarity to the Ideal Solution (TOPSIS) are used for the analysis of decision-making variables in a holistic way. AHP is incorporated into the model to determine the importance weights of the decision criteria, while the TOPSIS technique is used to obtain the performance ratings of decision alternatives.

The rest of the paper is structured as described in the next sentences. Following a brief literature review of decision-making analysis, a model for selecting the best ballast water treatment system is presented. Then the proposed model is demonstrated using a test case, with a sensitivity analysis to validate the findings, before the conclusion at the end.

2 Background to the Research

Bellman and Zadeh [3] surveyed decision-making problems using fuzzy sets, and initiated a Fuzzy Multiple Criteria Decision Making (FMCDM) methodology to resolve the lack of precision, in assigning importance weights of criteria and the ratings of alternatives [4, 5]. FMCDM has been applied in broad fields that include: the selection of strategic alliances partners for liner shipping [6]; safety assessment [7]; tool steel material selection [8]; assessment of climate change [9]; distribution centre location selection [10]; selection of a maintenance strategy for marine and offshore machinery operations [11]; and airline service quality evaluation [12].

A Multiple Criteria Decision Making (MCDM) problem can be defined as follows:

Let $A = \{A_i \text{ for } i = 1, 2, \ldots, m\}$ be a (finite) set of decision alternatives and G = $\{g_i \text{ for } j = 1,2,3, \ldots, n\}$ be a (finite) set of goals according to which the desirability of an action is judged. Determine the optimal alternative with the highest degree of desirability with respect to all relevant goals g_i [13].

Linguistic term sets used for describing each fundamental parameter are determined according to the situation of the case of interest [14]. However, some literature [15–17] shows that the number of linguistic terms ranging between four and seven labels, is commonly acceptable to represent risk factors in engineering risk analysis. In this study, five linguistic terms have been used to describe the evaluation criteria.

TOPSIS is a linear weighting technique which was first proposed in its crisp version by Chen and Hwang [18] with reference to Hwang and Yoon's work [19]. The technique was developed based on the concept that the chosen alternative should have the shortest distance from the positive ideal reference point (PIRP), and the farthest distance from the negative ideal reference point (NIRP) [20]. Assume that each attribute in the decision matrix takes either a monotonically increasing or monotonically decreasing utility; it will be easier to locate the positive ideal solution, which is a combination of all the best attribute values attainable, while the negative ideal solution is a combination of all the worse attribute values attainable [21].

TOPSIS has been proved to be one of the best methods in addressing the rank reversal issue; that is, the change in the ranking of alternatives when a non-optimal alternative is introduced [19]. Moreover, it has been proved to be insensitive to the number of alternatives and has its worst performance only in case of a very limited number of criteria. TOPSIS has been applied in varied and robust fields such as: evaluation and selection of initial training aircraft [5]; outsourcing of third party logistics service providers [19]; materials selection [22]; evaluation of competitive companies [23]; and the assessment of service quality in the airline industry [12].

Fuzzy-TOPSIS is a fuzzy extension of TOPSIS to efficiently handle the fuzziness of data to be applied in the decision-making process. A fuzzy approach to TOPSIS is often advantageous, because it assigns the relative importance of attributes using fuzzy numbers instead of precise numbers. Linguistic preferences can easily be converted to fuzzy numbers and TOPSIS allows the use of these fuzzy numbers in the calculation.

In order to apply a fuzzy TOPSIS to a MCDM problem, selection criteria have to be monotonic. Monotonic criteria could be classified either as benefits (B) or as costs (C). In fuzzy TOPSIS, the cost criteria are defined as the most desirable candidates scoring at the lowest, while the benefit criteria are described as the most desirable candidate scoring at the highest. Other advantages of the Fuzzy-TOPSIS technique include the fact that [19, 23, 24]:

The logic is rational and understandable.

- Computation processes are straightforward.
- The concept permits the pursuit of best alternatives for each criterion depicted in a simple mathematical form.
- It allows the straight linguistic definition of weights and ratings under each criterion, without the need of cumbersome pairwise comparisons and the risk of inconsistencies.
- The obtained weights of evaluation criteria are incorporated into the comparison procedures.

Given the stochastic nature of species assemblages, current inadequacy of historical data on non-indigenous invasive species (NIS) origin and dispersal mechanisms within the bio-geographical regions of the world, the fuzzy TOPSIS model has been proposed as an alternative technique for use in the analysis of ballast water treatment decision options. While the uncertainty issue is tackled by means of fuzzy logic, the application of TOPSIS makes it possible to investigate the distances of each decision option from the PIRP and NIRP. Moreover, the way linguistic ratings and weights are given is very straightforward.

The triangular fuzzy numbers are applied in the Fuzzy-TOPSIS used in this study. This is because it is intuitively easy for the decision-makers to use and calculate [25]. Secondly, modelling using triangular fuzzy numbers has proven to be an effective way for the formulation of the decision problem, where the information is subjective and imprecise [25].

Let $\tilde{A}$ and $\tilde{B}$ be two triangular fuzzy numbers denoted by the triplet (a_1, a_2, a_3) and (b_1, b_2, b_3) respectively. Then the basic fuzzy arithmetical operations on these two fuzzy numbers are defined as [26]:

$$\tilde{A}(+)\tilde{B} = (a_1, a_2, a_3)(+)(b_1, b_2, b_3) = (a_1 + b_1, a_2 + b_2, a_3 + b_3) \tag{1}$$

$$\tilde{A}(-)\tilde{B} = (a_1, a_2, a_3)(-)(b_1, b_2, b_3) = (a_1 - b_3, a_2 - b_2, a_3 - b_1) \tag{2}$$

$$\tilde{A}(\times)\tilde{B} = (a_1, a_2, a_3)(\times)(b_1, b_2, b_3) = (a_1 b_1, a_2 b_2, a_3 b_3) \tag{3}$$

$$\tilde{A}(\div)\tilde{B} = (a_1, a_2, a_3)(\div)(b_1, b_2, b_3) = \left(\frac{a_1}{b_3}, \frac{a_2}{b_2}, \frac{a_3}{b_1}\right) \tag{4}$$

The distance between fuzzy numbers $\tilde{A}$ and $\tilde{B}$ can be measured using the vertex method [27] and calculated using the following equation:

$$d(\tilde{A}, \tilde{B}) = \sqrt{\frac{1}{3}\left[(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2\right]} \tag{5}$$

3 A Proposed Model for Selecting the Best Ballast Water Treatment System

The proposed model and hierarchical structure describing the decision-making process of selecting the best ballast water treatment system is graphically illustrated in Fig. 1. The first stage is the identification of decision-making alternatives for ship-based ballast water treatment. The decision alternatives are literature-based and have been derived from the IMO Ballast Water Convention 2004 and the Lloyds

Report 2007 [1, 28]. The evaluation process is conducted by decision analysts based on their subjective knowledge and judgment.

The second stage in the model is the identification of the evaluation criteria for the identified prototype treatment technologies. In the third stage, AHP is applied to obtain the importance weights of the evaluation criteria. In the fourth stage, Fuzzy-TOPSIS is applied to obtain the performance ratings of the various decision alternatives.

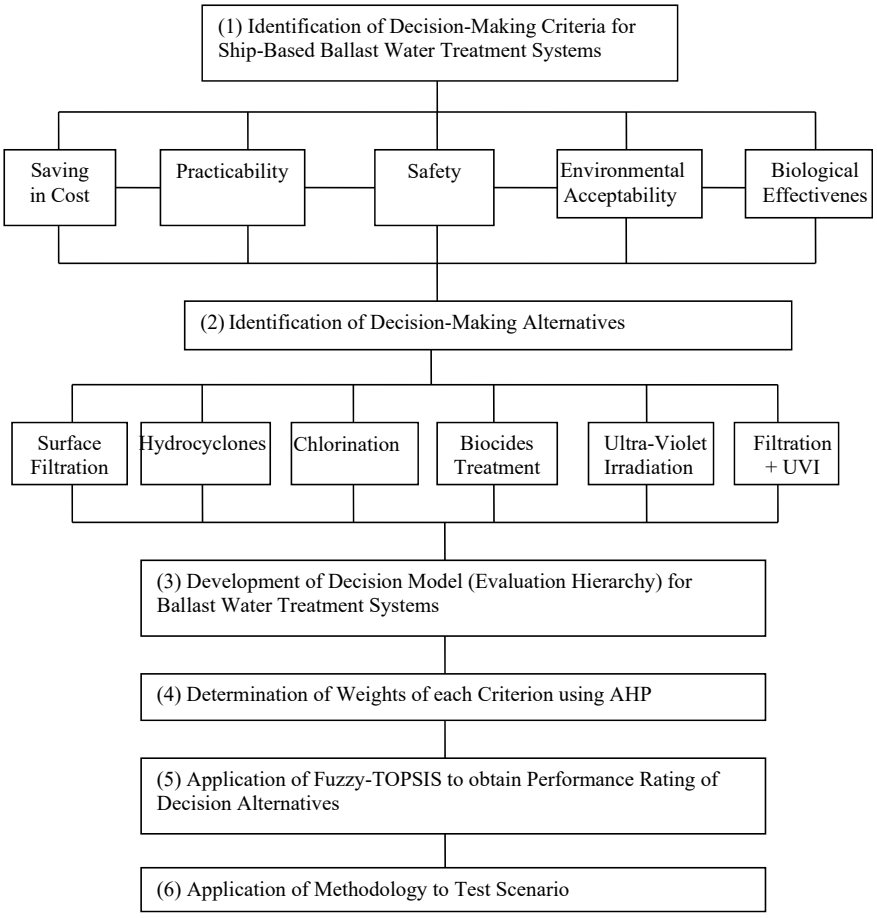


Fig. 1 Hierarchical model of decision making analysis

3.1 Identification of Evaluation Criteria

Five evaluation criteria have been identified for the evaluation of the decision alternatives. The criteria are based on the IMO guidelines for the development of prototype treatment technologies for on-board ballast water treatment [28]. They include:

- Saving in Cost (savings in expense of treatment equipment and operations).
- Practicability (ease of operating treatment equipment and interference with normal ship operations, as well as impact on the structural integrity of the ship).
- Safety (of crew, ship and cargo).
- Environmental Acceptability (not causing more or greater environmental impact than it solves).
- Biological Effectiveness (efficacy or effectiveness of removing or otherwise rendering inactive harmful non-indigenous invasive species (NIS) in ballast water).

3.2 Identification of Decision-Making Alternatives

Six decision-making alternatives have been identified and applied in this model, including surface filtration, hydro-cyclones, chlorination, biocides treatment, ultra-violet irradiation, and the combination of filtration and ultra-violet irradiation. The treatment systems have been selected from the generic ballast water treatment technologies [physical solid–liquid separation (primary treatment) and disinfection (secondary treatment)] recommended by the IMO for the global maritime industry [1].

3.3 Determination of Importance Weights of Decision Alternatives Using AHP

The next step in the methodology is the determination of importance weights of the five criteria described above, using the AHP approach involving a panel of domain experts. A consistency check is conducted to ensure that the pair-wise comparisons in the AHP are within the acceptable consistency. Experts may be revisited for their judgements, if the consistency of pair-wise comparisons is outside the required limit.

3.4 Application of Fuzzy-TOPSIS Approach to Obtain Performance Rating of Decision Alternatives

In this assessment process, all the variables are fuzzy variables represented by triangular fuzzy numbers. The process is conducted as follows.

3.4.1 Construction of Fuzzy Decision Matrix

A decision matrix is an $(m \times n)$ matrix in which $element_{ij}$ indicates the performance of alternative A_i when it is evaluated in terms of decision criterion C_j $(i = 1, 2, 3, \ldots, m; j = 1, 2, 3, \ldots, n)$ [7]. From this definition, it is implied that an MCDM problem with a given decision matrix is in essence a problem for a set of known alternatives and a set of known criteria [7].

Given m alternatives, n criteria and s decision analysts, a typical fuzzy MCDM problem can be represented using the following matrix [5, 19]:

$$R_k = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array} \\ \begin{bmatrix} \tilde{r}_{11} & \tilde{r}_{12} & \cdots & \tilde{r}_{1n} \\ \tilde{r}_{21} & \tilde{r}_{22} & \cdots & \tilde{r}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \tilde{r}_{m1} & \tilde{r}_{m2} & \cdots & \tilde{r}_{mn} \end{bmatrix} \end{array} \tag{6}$$

where, $A_1, A_2, \ldots, A_m$ represent the decision alternatives; $C_1, C_2, \ldots, C_n$ represent the evaluation criteria, and $\tilde{r}_{ij}$ represents the rating of the alternative A_i when examined in terms of criterion C_j evaluated by the s decision analysts.

3.4.2 Normalisation of Fuzzy Decision Matrix

The fuzzy data obtained in the decision matrix are normalised in order to eliminate the units of criteria scores, so that numerical comparisons associated with MCDM problems can be brought to the same universe of discourse. Normalisation has two main aims: for the comparison of heterogeneous criteria, and to ensure that all triangular fuzzy numbers range within the interval between 0 and 1 [5]. The normalised fuzzy-decision matrix is conducted using Eqs. 7 and 8 as follows:

If $\tilde{R}$ denotes the normalised fuzzy decision matrix, then

$$\tilde{R} = \left[\tilde{r}_{ij}\right]_{mxn}, \; i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n \tag{7}$$

$$\tilde{r}_{ij} = \left(\frac{a_{ij}}{c_j^+}, \frac{b_{ij}}{c_j^+}, \frac{c_{ij}}{c_j^+}\right) j \in B; \quad \tilde{r}_{ij} = \left(\frac{a_j^-}{c_{ij}}, \frac{a_j^-}{b_{ij}}, \frac{a_j}{a_{ij}}\right) j \in C \tag{8}$$

where

$$c_j^+ = \max_i c_{ij} \text{ when } \quad j \in B; \quad a_j^- = \min_i a_{ij} \text{ when } j \in C.$$

3.4.3 Construction of Weighted Normalised Fuzzy Decision Matrix

The process involves multiplying the importance weights of the criteria by their corresponding values in the normalised fuzzy decision matrix. Considering the different importance of each criterion, the weighted normalized fuzzy-decision matrix $\tilde{V}$ is constructed as:

$$\tilde{V} = \left[\tilde{v}_{ij}\right]_{m\times n} i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n \tag{9}$$

$$\tilde{v}_{ij} = \tilde{R}_{ij} \times \tilde{w}_j \tag{10}$$

where $\tilde{w}_j$ denotes the importance weight of criterion C_j.

3.4.4 Determination of the Fuzzy Positive Ideal Reference Point (FPIRP) and Fuzzy Negative Ideal Reference Point (FNIRP)

The FPIRP is obtained by identifying the best score in a criterion. Similarly, the worst score of a criterion is identified and recorded as the FNIRP. Against the background that all the triangular fuzzy numbers in $\tilde{V}$ are in the interval (0, 1), the FPIRP (A^+) (the benefit criterion) and FNIRP (A^-) (the cost criterion) are defined as follows [4]:

$$A^+ = (\tilde{v}_1^+, \tilde{v}_2^+, \rightleftarrows \ldots\ldots, \tilde{v}_n^+) \tag{11}$$

$$A^- = (\tilde{v}_1^-, \tilde{v}_2^-, \rightleftarrows \ldots\ldots, \tilde{v}_n^-) \tag{12}$$

where $\tilde{v}_j^+ = (1, 1, 1)$ and $\tilde{v}_j^- = (0, 0, 0)$, $j = 1, 2, \ldots, n$.

3.4.5 Calculation of Distances of Each Alternative to FPIRP and FNIRP

The distance of each alternative (treatment system) from the FPIRP and FNIRP with respect to each criterion is calculated as follows.

$$d_i^+ = \sum_{j=1}^{n} d\left(\tilde{v}_{ij}, \tilde{v}_j^+\right) \tag{13}$$

$$d_i^- = \sum_{j=1}^{n} d\left(\tilde{v}_{ij}, \tilde{v}_j^-\right) \tag{14}$$

where d_i^+ denotes the distance of alternative A_i from FPIRP, $d\left(\tilde{v}_{ij}, \tilde{v}_j^+\right)$ denotes the distance measurement between $\tilde{v}_{ij}$ and $\tilde{v}_j^+$; d_i^- is the distance of alternative A_i from FNIRP, and $d\left(\tilde{v}_{ij}, \tilde{v}_j^-\right)$ denotes the distance measurement between $\tilde{v}_{ij}$ and $\tilde{v}_j^-$.

3.4.6 Obtain the Closeness Coefficient and Ranking of Alternatives

The ranking of the alternatives can be determined after the Closeness Coefficient (CC_i) associated with A_i is obtained. This allows the decision maker to choose the most rational alternative. CC_i can be calculated by:

$$CC_i = \frac{d_i^-}{d_i^+ + d_i^-} i = 1, 2, \ldots, m \tag{15}$$

where CC_i is equal to 0 if and only if $d_i{}^- = 0$ or $A_i = A^-$. $CC_i = 1$ when $d_i{}^+ = 0$ or $A_i = A^+$. As a result, the best alternative is the one with the value of CC_i closest to 1.

4 Application of Methodology to a Test Scenario

The proposed model will be demonstrated in a decision analysis of selecting on-board ballast water treatment technologies. In this study, five qualified and experienced experts have been identified to conduct the analysis. The analysts are assigned equal ratings and the analysis will be conducted through brainstorming based on their knowledge and experience.

The weight values for the evaluation criteria are obtained as follows using the AHP approach [29]:

Saving in Cost = 0.068
Practicability = 0.171
Safety = 0.392
Environmental Acceptability = 0.237
Biological Effectiveness = 0.132

The importance weight distributions for the decision-making criteria show that the criterion “Safety” recorded the highest weight (0.392), whereas the lowest weight

Table 1 TOPSIS decision alternatives and evaluation criteria

Decision alternatives		Evaluation criteria	
A_1	Surface filtration	C_1	Saving in cost
A_2	Hydrocyclones	C_2	Practicability
A_3	Chlorination	C_3	Safety
A_4	Biocides	C_4	Environmental acceptability
A_5	UV irradiation	C_5	Biological effectiveness
A_6	Filtration + UV irradiation		

(0.068) is associated with the criterion "Saving in Cost". These importance weights will be applied in the next stage of this study, to establish the fuzzy performance ratings of the evaluation criteria.

The six decision alternatives and five evaluation criteria (Table 1) will be used to develop the fuzzy decision matrix.

4.1 Construction of a Fuzzy-TOPSIS Decision Matrix

The membership functions of the linguistic variables, and scales developed for the measurement of the importance of the evaluation criteria, are shown in Table 2. A Fuzzy-TOPSIS decision matrix is then constructed as shown in Table 3.

Table 2 Fuzzy-linguistic scales for measuring performance of evaluation criteria

Linguistic variable	Corresponding triangular fuzzy number
Very poor	(0, 1, 3)
Poor	(1, 3, 5)
Average	(3, 5, 7)
Good	(5, 7, 9)
Very good	(7, 9, 10)

Table 3 Fuzzy TOPSIS decision matrix

	C_1	C_2	C_3	C_4	C_5
A_1	5, 7, 9	7, 9, 10	5, 7, 9	7, 9, 10	5, 7, 9
A_2	5, 7, 9	5, 7, 9	5, 7, 9	7, 9, 10	5, 7, 9
A_3	3, 5, 7	5, 7, 9	5, 7, 9	3, 5, 7	5, 7, 9
A_4	3, 5, 7	5, 7, 9	3, 5, 7	1, 3, 5	5, 7, 9
A_5	5, 7, 9	5, 7, 9	3, 5, 7	5, 7, 9	5, 7, 9
A_6	5, 7, 9	7, 9, 10	7, 9, 10	7, 9, 10	7, 9, 10

Table 4 Fuzzy TOPSIS normalized decision matrix

	C_1	C_2	C_3	C_4	C_5
A_1	0.5555, 0.7777, 1.0000	0.7000, 0.9000, 1.0000	0.5000, 0.7000, 0.9000	0.7000, 0.9000, 1.0000	0.5000, 0.7000, 0.9000
A_2	0.5555, 0.7777, 1.0000	0.5000, 0.7000, 0.9000	0.5000, 0.7000, 0.9000	0.5000, 0.7000, 0.9000	0.5000, 0.7000, 0.9000
A_3	0.3333, 0.5555, 0.7777	0.5000, 0.7000, 0.9000	0.5000, 0.7000, 0.9000	0.3000, 0.5000, 0.7000	0.5000, 0.7000, 0.9000
A_4	0.3333, 0.5555, 0.7777	0.5000, 0.7000, 0.9000	0.3000, 0.5000, 0.7000	0.1000, 0.3000, 0.5000	0.5000, 0.7000, 0.9000
A_5	0.5555, 0.7777, 1.0000	0.5000, 0.7000, 0.9000	0.3000, 0.5000, 0.7000	0.5000, 0.7000, 0.9000	0.5000, 0.7000, 0.9000
A_6	0.5555, 0.7777, 1.0000	0.7000, 0.9000, 1.0000	0.7000, 0.9000, 1.0000	0.7000, 0.9000, 1.0000	0.7000, 0.9000, 1.0000

4.2 Normalisation of Fuzzy Decision Matrix

The normalized fuzzy decision matrix is constructed using Eqs. 7 and 8. The results are shown in Table 4.

4.3 Construction of Weighted Normalised Fuzzy-Decision Matrix

The weighted normalized decision matrix is constructed by applying Eqs. 9 and 10. The normalized triangular fuzzy numbers are obtained as shown in Table 5. For example, the weighted normalized fuzzy number for A_3 with respect to C_2 is obtained as follows.

$$(0.500,\ 0.700,\ 0.900) \times 0.171 = (0.086,\ 0.120,\ 0.154)$$

4.4 Determination of the Fuzzy Positive Ideal Reference Point (FPIRP) and Fuzzy Negative Ideal Reference Point (FNIRP)

D_i^+ and d_i^- are obtained using Eqs. 13 and 14. For example, d_1^+ and d_1^- are obtained as follows:

Table 5 Weighted Normalised Decision Matrix of the Six Ballast Water Treatment Systems

	C_1	C_2	C_3	C_4	C_5
A_1	0.038, 0.053, 0.068	0.119, 0.153, 0.171	0.196, 0.274, 0.352	0.165, 0.213, 0.237	0.066, 0.092, 0.118
A_2	0.038, 0.053, 0.068	0.086, 0.120, 0.154	0.196, 0.274, 0.353	0.166, 0.213, 0.237	0.066, 0.092, 0.119
A_3	0.023, 0.038, 0.053	0.086, 0.120, 0.154	0.196, 0.274, 0.353	0.071, 0.119, 0.166	0.066, 0.092, 0.119
A_4	0.023, 0.038, 0.053	0.086, 0.120, 0.154	0.118, 0.196, 0.274	0.024, 0.071, 0.119	0.066, 0.092, 0.119
A_5	0.038, 0.053, 0.068	0.086, 0.120, 0.154	0.118, 0.196, 0.274	0.119, 0.166, 0.213	0.066, 0.092, 0.119
A_6	0.038, 0.053, 0.068	0.120, 0.154, 0.171	0.274, 0.353, 0.392	0.166, 0.213, 0.237	0.092, 0.119, 0.132

$$
\begin{aligned}
d_1^+ = &\sqrt{\frac{1}{3}\left[(0.0378-1)^2+(0.0539-1)^2+(0.0680-1)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.1197-1)^2+(0.1539-1)^2+(0.1710-1)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.1960-1)^2+(0.2744-1)^2+(0.3528-1)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.15689-1)^2+(0.2744-1)^2+(0.3528-1)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.0660-1)^2+(0.0924-1)^2+(0.1188-1)^2\right]} = 4.231
\end{aligned}
$$

$$
\begin{aligned}
d_1^- = &\sqrt{\frac{1}{3}\left[(0.0378-0)^2+(0.0529-0)^2+(0.068-0)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.1197-0)^2+(0.1539-0)^2+(0.1710-0)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.1960-0)^2+(0.2744-0)^2+(0.3528-0)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.1659-0)^2+(0.2133-0)^2+(0.2370-0)^2\right]} \\
&+\sqrt{\frac{1}{3}\left[(0.066-0)^2+(0.0924-0)^2+(0.1188-0)^2\right]} = 0.788
\end{aligned}
$$

The distances of the other decision alternatives to the FRIRP and ENIRP were determined in the same way and the results are described in Table 6.

Table 6 Results of Fuzzy TOPSIS analysis

	Decision making criterion	d^+	d^-	Closeness coefficient	Ranking
A_1	Surface filtration	4.231	0.788	0.157	2
A_2	Hydrocyclones	4.299	0.724	0.144	3
A_3	Chlorination	4.362	0.663	0.132	4
A_4	Biocides	4.487	0.545	0.108	6
A_5	UV irradiation	4.377	0.649	0.129	5
A_6	Filtration + UV irradiation	4.142	0.870	0.174	1

4.5 Obtain Closeness Co-efficient and Ranking of Alternatives

The treatment system with a larger CC value is more desirable. The calculation of the CC value has been described below using A_1 as an example.

$$d_1^+ = 4.231 \quad d_1^- = 0.788 \quad CC_1 = \frac{0.788}{4.231 + 0.788} = 0.157$$

By applying the same method, the Closeness Coefficient values of attributes A_2–A_6 are obtained as shown in Table 6.

5 Results and Validation of Model

From the result of the Fuzzy-TOPSIS analysis (Table 6), it can be seen that the highest CC value (0.174) is associated with A_6 (Filtration + UV Irradiation). The lowest CC value (0.108) is associated with A_4 (Biocides). The result also shows that A_2 is ranked third with a CC value of 0.144. A_3 is ranked fourth having returned a CC value of 0.132, while A_5 is placed fifth in the ranking with a CC value of 0.129.

The result also shows that the CC values of the six decision alternatives are marginally separated. This suggests the degree of reasonableness and relative closeness of the systems for the treatment of ships' ballast water. Based on the output values obtained in this analysis, the ranking (in order of preference) of the six decision alternatives in descending order is: $A_6 > A_1 > A_2 > A_3 > A_5 > A_4$.

In order to validate and test the robustness of this model, a sensitivity analysis is conducted. The analysis is necessary in order to test the suitability and sensitivity of the model for decision analysis of prototype ballast water treatment technologies (as decision alternatives). The analysis is conducted under eight conditions as tabulated in Table 7.

The first step in the sensitivity analysis process, involves an increment of the main values of the positive and negative reference points (d^+ and d^-), of each decision

Table 7 Conditions for changing output values by percentages

Condition	Percentage
1	Increase d^+ by 5%
2	Increase d^- by 5%
3	Decrease d^+ by 5%
4	Decrease d^- by 5%
5	Increase d^+ by 20%
6	Increase d^- by 20%
7	Decrease d^+ by 20%
8	Decrease d^- by 20%

alternative by 5 and 20%. The next step is to decrease the same values separately by 5 and 20%.

From the results of the sensitivity analysis (Table 8), it can be seen that the ranking order of the six decision alternatives maintain a consistency when d^+ and d^- of each alternative are increased by 5 and 20%. Such a ranking order also maintains a consistency when d^+ and d^- of each alternative are decreased by 5 and 20%. The results also show that the Closeness Coefficient values of A_1–A_6 consistently increase in Conditions 1, 2, 5 and 6. The Closeness Coefficient values of A_1–A_6 consistently decrease in Conditions 3, 4, 7 and 8. This pattern in the results is to be expected. The model is reasonable and capable of being applied in the analysis of ballast water decision-making alternatives.

6 Conclusion

This model was developed by taking into consideration the legislative requirements of Regulation D2–D4 of the IMO Ballast Water Convention 2007, as well as the positive contributions of the scientific and technological communities in developing prototype ballast water treatment systems. It is pertinent to state that the inadequacy of data and/or stochastic nature of species assemblages within the global bio-geographical regions pose a great threat to the attainment of the IMO Standards and the utilization of any developed treatment systems for the management of NIS.

It therefore remains uncertain that, a chosen treatment system would be safe, practicable, cost effective, environmentally acceptable, or biologically effective in minimizing the survivability of ballast tank based NIS. This uncertainty can result in the selection of an inappropriate treatment system for the wrong ship type and/or wrong voyage route, thus leading to severe environmental and/or financial consequences.

Powerful MCDM methodologies (AHP and TOPSIS) were applied in this generic model, to solve inherent decision-making problems that could be encountered during

Table 8 Results of sensitivity analysis

		A_1			A_2			A_3		
Condition		d^+	d^-	CC_i	d^+	d^-	CC_i	d^+	d^-	CC_i
Main		4.231	0.788	0.157	4.299	0.724	0.144	4.362	0.663	0.132
1	Increase d^+ by 5%	4.442	0.788	0.151	4.514	0.724	0.138	4.580	0.663	0.126
2	Increase d^- by 5%	4.231	0.827	0.164	4.299	0.688	0.138	4.362	0.696	0.138
3	Decrease d^+ by 5%	4.019	0.788	0.164	4.084	0.724	0.151	4.144	0.663	0.138
4	Decrease d^- by 5%	4.231	0.749	0.150	4.299	0.688	0.138	4.362	0.630	0.126
5	Increase d^+ by 20%	5.077	0.788	0.134	4.444	0.724	0.140	5.234	0.663	0.112
6	Increase d^- by 20%	4.231	0.946	0.183	4.299	0.869	0.144	4.362	0.796	0.154
7	Decrease d^+ by 20%	3.385	0.788	0.189	3.439	0.724	0.174	3.490	0.663	0.160
8	Decrease d^- by 20%	4.231	0.630	0.130	4.299	0.579	0.119	4.362	0.530	0.108
Main		A_4			A_5			A_6		
1	Increase d^+ by 5%	4.711	0.545	0.104	4.596	0.649	0.124	4.349	0.870	0.167
2	Increase d^- by 5%	4.487	0.572	0.113	4.377	0.681	0.135	4.142	0.914	0.181
3	Decrease d^+ by 5%	4.263	0.545	113	4.158	0.649	0.135	3.935	0.870	0.181
4	Decrease d^- by 5%	4.487	0.518	0.103	4.377	0.617	0.124	4.142	0.827	0.166
5	Increase d^+ by 20%	5.384	0.545	0.092	5.252	0.649	0.110	4.970	0.870	0.149
6	Increase d^- by 20%	4.487	0.654	0.127	4.377	0.779	0.151	4.142	1.044	0.201
7	Decrease d^+ by 20%	3.599	0.545	0.132	3.502	0.649	0.156	3.314	0.870	0.208
8	Decrease d^- by 20%	4.487	0.436	0.089	4.377	0.519	0.106	4.142	0.696	0.140

the selection process of a ballast water treatment technology under a fuzzy environment. These methodologies have been applied in different specialized fields as stated earlier and found to be effective.

The model developed in this study is by no means conclusive. It is subject to further modification given the acquisition of new data, or current status before its utilization by end-users in the industry. Lastly, a sensitivity analysis was conducted to

partially validate the developed model, and establish its ability to respond to changes in input variables.

References

1. Lloyd's Register (2007) Ballast water treatment technology: current status. London
2. Cheng CH, Lin Y (2002) Evaluating the best main battle tank using fuzzy decision theory with linguistic criteria evaluation. Eur J Oper Res 142(1):174–186
3. Bellman RE, Zadeh LA (1970) Decision-making in a fuzzy environment. Manage Sci 17(4):B-141
4. Chen CB, Klein CM (1997) An efficient approach to solving fuzzy MADM problems. Fuzzy Sets Syst 88(1):51–67
5. Wang TC, Chang TH (2007) Application of TOPSIS in evaluating initial training aircraft under a fuzzy environment. Expert Syst Appl 33(4):870–880
6. Ding JF, Liang GS (2005) Using fuzzy MCDM to select partners of strategic alliances for liner shipping. Inf Sci 173(1–3):197–225
7. Schinas O (2007, September) Examining the use and application of multi-criteria decision making techniques in safety assessment. In: International symposium on maritime safety, security and environmental protection. Athens, pp 20–21
8. Shyi-Ming C (1997) A new method for tool steel materials selection under fuzzy environment. Fuzzy Sets Syst 92(3):265–274
9. Bell ML, Hobbs BF, Ellis H (2003) The use of multi-criteria decision-making methods in the integrated assessment of climate change: implications for IA practitioners. Socioecon Plann Sci 37(4):289–316
10. Chen CT (2001) A fuzzy approach to select the location of the distribution center. Fuzzy Sets Syst 118(1):65–73
11. Asuquo MP, Wang J, Zhang L, Phylip-Jones G (2019) Application of a multiple attribute group decision making (MAGDM) model for selecting appropriate maintenance strategy for marine and offshore machinery operations. Ocean Eng 179:246–260
12. Tsaur SH, Chang TY, Yen CH (2002) The evaluation of airline service quality by fuzzy MCDM. Tour Manage 23(2):107–115
13. Zimmermann HJ (1991) Fuzzy set theory and its applications, 2nd edn. Kluwer Academic, Boston
14. Liu J, Yang JB, Wang J, Sii HS, Wang YM (2004) Fuzzy rule-based evidential reasoning approach for safety analysis. Int J Gen Syst 33(2–3):183–204
15. Karwowski W, Mital A (1986) Potential applications of fuzzy sets in industrial safety engineering. Fuzzy Sets Syst 19(2):105–120
16. Bowles JB, Peláez CE (1995) Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis. Reliab Eng Syst Saf 50(2):203–213
17. Wang J (1997) A subjective methodology for safety analysis of safety requirements specifications. IEEE Trans Fuzzy Syst 5(3):418–430
18. Chen SJ, Hwang CL (1992) Fuzzy multiple attribute decision making methods. In: Fuzzy multiple attribute decision making. Springer, Berlin, pp 289–486
19. Bottani E, Rizzi A (2006) A fuzzy TOPSIS methodology to support outsourcing of logistics services. Supply Chain Manage Int J 11(4):294–308
20. Tzeng GH, Huang JJ (2011) Multiple attribute decision making: methods and applications. CRC Press
21. Yoon K, Hwang C (1995) Multi-attribute decision making: an introduction. Sage Publications, London
22. Jee DH, Kang KJ (2000) A method for optimal material selection aided with decision making theory. Mater Des 21(3):199–206

23. Deng H, Yeh CH, Willis RJ (2000) Inter-company comparison using modified TOPSIS with objective weights. Comput Oper Res 27(10):963–973
24. Olson DL (2004) Comparison of weights in TOPSIS models. Math Comput Model 40(7–8):721–727
25. Dağdeviren M, Yavuz S, Kılınç N (2009) Weapon selection using the AHP and TOPSIS methods under fuzzy environment. Expert Syst Appl 36(4):8143–8151
26. Dubois D, Prade H (1997) Recent models of uncertainty and imprecision as a basis for decision theory: toward less normative frameworks. In: Hollnagel E, Mancini G, Woods DD (eds) Intelligent decision support in process environment. Springer, New York
27. Chen CT (2000) Extensions of the TOPSIS for group decision-making under fuzzy environment. Fuzzy Sets Syst 114(1):1–9
28. IMO (2004) International convention for the control and management of ships' ballast water and sediments. International Maritime Organisation, London. Can be accessed at: http://globallast.imo.org/index.asp?page=mepc.htm&menu=true
29. Pam ED (2010) Risk-based framework for ballast water safety management, PhD thesis, Liverpool John Moores University, UK

A Two-Phase Sampling Approach for Reliability-Based Optimization in Structural Engineering

Danko J. Jerez, Hector A. Jensen, and Michael Beer

Abstract This work presents a two-phase sampling approach to address reliability-based optimization problems in structural engineering. The constrained optimization problem is converted into a sampling problem, which is then solved using Markov chain Monte Carlo methods. First, an exploration phase generates uniformly distributed feasible designs. Thereafter, an exploitation phase is carried out to obtain a set of close-to-optimal designs. The approach is general in the sense that it is not limited to a particular type of system behavior and, in addition, it can handle constrained and unconstrained formulations as well as discrete–continuous design spaces. Three numerical examples involving structural dynamical systems under stochastic excitation are presented to illustrate the capabilities of the approach.

Keywords Reliability-based optimization · Structural engineering · First-passage probability · Metamodel · Stochastic search

1 Introduction

Structural engineering practice is inherently related to the design of safe and cost-efficient systems to satisfy private and public needs. Optimization techniques have proved instrumental to this end, whereby suitable design solutions are typically identified by minimizing a cost function subject to certain constraints [1, 2]. Since the

D. J. Jerez (✉) · M. Beer
Institute for Risk and Reliability, Leibniz Universität Hannover, Hannover, Germany
e-mail: danko.jerez@irz.uni-hannover.de

H. A. Jensen
Department of Civil Engineering, Federico Santa Maria Technical University, Valparaíso, Chile

M. Beer
International Joint Research Center for Resilient Infrastructure and International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Shanghai, China

Institute for Risk and Uncertainty and School of Engineering, University of Liverpool, Liverpool, UK

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_2

systems of interest are unavoidably exposed to external actions and deterioration processes that are difficult to predict, the treatment of uncertainties is a key aspect to obtain meaningful optimization results. In this regard, reliability-based optimization (RBO) offers a rational and theoretically sound framework to incorporate the interaction between uncertainties and design requirements into decision-making processes [3–5]. In this setting, system performance metrics are explicitly included in the objective and/or constraint functions by means of reliability measures.

Structural systems of practical interest are characterized by their exposure to environmental actions, a relatively large scale, and a complex behavior. In this context, the prediction of the system response often relies on complex computational procedures involving, e.g., the numerical solution of nonlinear equations with multiple unknowns. Further, their probabilistic characterization using, for instance, random fields or stochastic processes requires a relatively large number of random variables, which leads to high-dimensional reliability integrals [6]. These features make reliability assessment a challenging task, which is usually addressed using stochastic simulation [7]. This, in turn, brings challenges to the solution of RBO problems associated with the computational cost, inherent variability, and sensitivity evaluation of the reliability estimates [8].

Several approaches have been reported to address RBO problems involving high-dimensional probability integrals. In general, they can be classified in three groups based on the adopted search strategy [8], namely, sequential optimization approaches, stochastic search-based techniques, and schemes based on augmented reliability spaces. Although the most suitable optimization approach depends on the problem characteristics, the use of stochastic search-based techniques can be regarded as a general and flexible strategy. These methods do not need sensitivity measures and are not restricted to a specific class of systems. However, they tend to be computationally more intensive than sequential strategies (see, e.g., [9, 10]) or augmented reliability formulations (see, e.g., [11, 12]).

This work presents a stochastic search-based approach for the RBO of structural engineering systems including high-dimensional reliability integrals. The method is based on a two-phase sampling framework [13–16]. An exploration phase is first carried out to obtain feasible designs, which are then used in an exploitation phase that ultimately yields a set of close-to-optimal designs. The method can handle unconstrained, constrained, discrete and continuous formulations [16]. In addition, a suitable metamodel is implemented for improved numerical efficiency [14, 17]. Three examples involving a class of structural systems, namely, structural dynamical systems under stochastic excitation, are presented. Overall, the method represents a potentially useful tool to address a practical class of RBO problems in engineering applications.

2 Reliability-Based Optimization

2.1 Formulation

The class of problems of interest can be stated as

$$\begin{aligned} \min_{\boldsymbol{x}} \; & f(\boldsymbol{x}) \\ \text{s.t.} \; & r_j(\boldsymbol{x}) \leq 0, \; j = 1, \ldots, n_r \\ & g_k(\boldsymbol{x}) \leq 0, \; k = 1, \ldots, n_g \\ & \boldsymbol{x} \in \boldsymbol{X} \end{aligned} \tag{1}$$

where the vector $\boldsymbol{x} \in \boldsymbol{X} \subset \mathbb{R}^{n_x}$ comprises the n_x design variables (continuous and/or discrete), $f(\boldsymbol{x})$ is a general objective function, $r_j(\boldsymbol{x}) \leq 0$, $j = 1, \ldots, n_r$, represent n_r constraints in terms of system reliability measures, and $g_k(\boldsymbol{x}) \leq 0$, $k = 1, \ldots, n_g$, are n_g standard constraints. The vector of design variables is expressed as $\boldsymbol{x}^T = \langle \boldsymbol{x}_c^T, \boldsymbol{x}_d^T \rangle$ with $\boldsymbol{x}_c \in \boldsymbol{X}_c \subset \mathbb{R}^{n_c}$ and $\boldsymbol{x}_d \in \boldsymbol{X}_d \subset \mathbb{R}^{n_c}$ containing, respectively, the n_c continuous and n_d discrete design variables. In this formulation, the set $\boldsymbol{X} = \boldsymbol{X}_c \times \boldsymbol{X}_d$ characterizes explicit constraints on the design variables. For the continuous components, side constraints are imposed as

$$\boldsymbol{X}_c = \left\{ \boldsymbol{x}_c \in \mathbb{R}^{n_c} : x_{ci}^L \leq x_{ci} \leq x_{ci}^U, i = 1, \ldots, n_c \right\} \tag{2}$$

where x_{ci}^L and x_{ci}^U are the lower and upper bounds of the ith continuous design variable, respectively. In addition, the side constraints for the discrete design variables are given by

$$\boldsymbol{X}_d = \left\{ \boldsymbol{x}_d \in \mathbb{R}^{n_d} : x_{di} \in \boldsymbol{X}_{di}, i = 1, \ldots, n_d \right\} \tag{3}$$

where the set $\boldsymbol{X}_{di} = \left\{ x_{di(m)}, m = 1, \ldots, n_{di} \right\}$ contains the n_{di} allowable values for the ith discrete component of the design vector. For convenience, it is assumed that these values are sorted in an ascending order.

In the previous setting, the objective function $f(\boldsymbol{x})$ can be related to, e.g., initial construction costs, life-cycle costs, or structural performance measures. Further, the standard constraints $g_k(\boldsymbol{x}) \leq 0$, $k = 1, \ldots, n_g$, are associated with design requirements such as material availability, budget restrictions, etc., that do not involve system reliability measures. Hence, it is assumed that the functions $g_k(\boldsymbol{x}), k = 1, \ldots, n_g$, are relatively inexpensive to compute. In addition, the reliability constraints represent design conditions expressed in terms of failure probabilities as

$$r_j(\boldsymbol{x}) = P_{F_j}(\boldsymbol{x}) - P_{F_j}^* \leq 0, \; j = 1, \ldots, n_r \tag{4}$$

where $P_{F_j}(\boldsymbol{x})$ is the probability of failure event F_j evaluated at design $\boldsymbol{x}$, and $P_{F_j}^*$ is the corresponding maximum allowable value. The failure events can be defined,

e.g., in terms of serviceability conditions, users' comfort requirements, and partial or total collapse. It is noted that, according to this formulation, failure probability measures can be involved in the definition of the objective function and/or reliability constraints. Thus, the optimization problem stated in Eq. (1) is quite general in the sense that it allows the treatment of several RBO formulations. In this regard, indicative applications include the design of wind-excited buildings [18], structural topology optimization [19], and energy harvester optimization [20].

2.2 First-Passage Probabilities

For a general class of complex engineering systems, a suitable reliability measure corresponds to the so-called first-passage probability [21]. This measure quantifies the likelihood of performance requirements not being satisfied at any instant of a reference period. In this framework, consider a vector of basic random variables $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ following the multivariate probability density function $q(\boldsymbol{\theta}|\boldsymbol{x})$, i.e., $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\boldsymbol{x})$. This distribution can depend, in principle, on the vector of design variables $\boldsymbol{x}$. If that is not the case, the random variables are simply distributed as $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$. In general, the vector $\boldsymbol{\theta}$ characterizes the uncertainty in the system properties as well as in the external actions over the system. Then, a first-passage failure event F can be defined as $F = \{d(\boldsymbol{x}, \boldsymbol{\theta}) > 1\}$ with normalized demand function $d(\boldsymbol{x}, \boldsymbol{\theta})$ given by

$$d(\boldsymbol{x}, \boldsymbol{\theta}) = \max_{t \in [0, t_T]} \max_{\ell = 1, \ldots, n_\eta} \frac{\eta_\ell(t, \boldsymbol{x}, \boldsymbol{\theta})}{\eta_\ell^*} \tag{5}$$

where $\eta_\ell(t, \boldsymbol{x}, \boldsymbol{\theta})$, $\ell = 1, \ldots, n_\eta$, are the system response functions of interest with corresponding thresholds $\eta_\ell^* > 0$, and t_T is the reference period. In general, these functions depend on time, the design variables $\boldsymbol{x}$, and the basic random variables $\boldsymbol{\theta}$. Hence, from the previous description, failure is defined when any response of interest exceeds its prescribed maximum allowable value at any instant of a reference period. Then, the corresponding first-passage probability can be written as

$$P_F(\boldsymbol{x}) = \int_{d(\boldsymbol{x}, \boldsymbol{\theta}) > 1} q(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta} \tag{6}$$

For most complex engineering systems, the vector of random variables $\boldsymbol{\theta}$ is high-dimensional and, in addition, the responses of interest are only available through involved black-box models. As a result, the evaluation of the previous integral is quite challenging, and it is usually carried out using stochastic simulation methods [22]. As previously pointed out, this makes the solution of RBO problems a challenging task due to the computational cost, noisy behavior and involved sensitivity estimation of failure probability functions [8].

3 Two-Phase Sampling Approach

3.1 Underlying Idea

Following the ideas of simulated annealing [23], the solution of the optimization problem in Eq. (1) can be equivalently formulated as the generation of designs that follow an appropriate probabilistic distribution. Such a formulation stems from the concept of canonical distribution in statistical mechanics [24] and the fact that finding the minimum of $f(\boldsymbol{x})$ is equivalent to maximizing the function $\exp(-f(\boldsymbol{x})/T)$ for any $T > 0$ [23]. Consider the auxiliary distribution

$$p(\boldsymbol{x}|T) \propto U_{\overline{\boldsymbol{X}}}(\boldsymbol{x}) \exp\left(-\frac{f(\boldsymbol{x})}{T}\right) \tag{7}$$

where $T > 0$ is the temperature parameter and $U_{\overline{\boldsymbol{X}}}(\boldsymbol{x})$ is a uniform distribution over the feasible set $\overline{\boldsymbol{X}}$, which is defined as

$$\overline{\boldsymbol{X}} = \left\{\boldsymbol{x} \in \boldsymbol{X} : r_j(\boldsymbol{x}) \leq 0,\ j = 1, \ldots, n_r \wedge g_k(\boldsymbol{x}) \leq 0, k = 1, \ldots, n_g\right\} \tag{8}$$

In Eq. (7), the parameter T affects the spread of the distribution $p(\boldsymbol{x}|T)$. On the one hand, increasing the value of T leads to flatter distributions. In the limit case in which $T \to \infty$, the auxiliary distribution becomes uniform over the feasible set, i.e., $\lim_{T\to\infty} p(\boldsymbol{x}|T) = U_{\overline{\boldsymbol{X}}}(\boldsymbol{x})$. On the other hand, for smaller values of T the distribution in Eq. (7) becomes increasingly concentrated around the feasible designs that minimize $f(\boldsymbol{x})$. In fact, when $T \to 0$ the probability mass is uniformly distributed over the optimal solution set $\boldsymbol{X}_f^*$, that is, $\lim_{T\to 0} p(\boldsymbol{x}|T) = U_{\boldsymbol{X}_f^*}(\boldsymbol{x})$. Thus, by generating samples (designs) that follow $p(\boldsymbol{x}|T)$, $T \to 0$, the optimal solution set corresponding to Eq. (1) can be explored. In other words, the solution of the RBO problem can be restated as the generation of samples according to the target distribution $\lim_{T\to 0} p(\boldsymbol{x}|T)$. It is noted that, in a Bayesian framework, the target distribution can be also interpreted as a posterior distribution where $U_{\overline{\boldsymbol{X}}}(\boldsymbol{x})$ plays the role of the prior distribution and $\lim_{T\to 0} \exp(-f(\boldsymbol{x})/T)$ of the (unnormalized) likelihood function [16].

3.2 Sequence of Intermediate Distributions

The straightforward generation of samples following $\lim_{T\to 0} p(\boldsymbol{x}|T)$ with, e.g., direct Monte Carlo simulation is generally unfeasible. To circumvent this issue, a sequential strategy is adopted in this work [25–28]. Consider the sequence of non-normalized intermediate distributions

$$
\begin{aligned}
&p_0(\boldsymbol{x}) = U_{\overline{\boldsymbol{X}}}(\boldsymbol{x}) \\
&p_j(\boldsymbol{x}) \propto U_{\overline{\boldsymbol{X}}}(\boldsymbol{x}) \exp\left(-\tfrac{f(\boldsymbol{x})}{T_j}\right),\ j = 1, 2, \ldots
\end{aligned}
\tag{9}
$$

where $\infty = T_0 > T_1 > \ldots > T_j > \ldots$ is a sequence of monotonically decreasing temperature parameters, with $T_j \to 0$ as $j \to \infty$, which are adaptively chosen to achieve a smooth transition between distributions. Such a strategy has been adopted to address several applications including, e.g., Bayesian model updating [25, 26], structural optimization [27, 28], and structural reliability assessment [29].

Based on the previous setting, it is seen that the initial distribution is uniform over the feasible set, whereas the next distributions in the sequence become increasingly concentrated around the optimal solution set as the temperature parameter decreases. Thus, the main idea is to generate samples (designs) in a sequential manner. In the initial stage ($j = 0$), samples uniformly distributed over the feasible set are obtained. Then, during stage $j = 1, 2, \ldots$, samples following the distribution $p_j(\boldsymbol{x})$ are drawn based on the samples from the previous stage. The transitional Markov chain Monte Carlo (TMCMC) method [25] is implemented for the sample generation process. Finally, once a certain stopping criterion is verified, the final designs represent a set of close-to-optimal solutions that follow an approximately uniform distribution over the optimal solution set.

The previous formulation requires an initial set of samples uniformly distributed over the feasible set $\overline{\boldsymbol{X}}$, which is usually difficult to obtain in a direct manner for practical cases. Thus, a two-phase sampling approach is adopted in this work [14, 16]. First, an *exploration phase* is carried out to obtain uniformly distributed designs in the feasible set. Then, these designs are used as the initial population of an *exploitation phase* which ultimately yields a set of close-to-optimal solutions.

3.3 Exploration Phase

To obtain designs following $U_{\overline{\boldsymbol{X}}}(\boldsymbol{x})$, consider the auxiliary optimization problem

$$
\begin{aligned}
&\min_{\boldsymbol{x}}\ h(\boldsymbol{x}) = \max\left\{0,\ \max_{j=1,\ldots,n_r} r_j(\boldsymbol{x}),\ \max_{k=1,\ldots,n_g} g_k(\boldsymbol{x})\right\} \\
&\text{s.t}\quad \boldsymbol{x} \in \boldsymbol{X}
\end{aligned}
\tag{10}
$$

From the previous definition, the minimum value of the auxiliary objective function $h(\boldsymbol{x})$ is equal to zero with corresponding optimal solution set [14]

$$
\boldsymbol{X}_h^* = \left\{\boldsymbol{x} \in \boldsymbol{X} : r_j(\boldsymbol{x}) \leq 0,\ j = 1, \ldots, n_r \wedge g_k(\boldsymbol{x}) \leq 0, k = 1, \ldots, n_g\right\} \tag{11}
$$

Thus, the optimal solution set in Eq. (11) is equal to the feasible design set in Eq. (8), i.e., $\boldsymbol{X}_h^* = \overline{\boldsymbol{X}}$. In addition, the auxiliary optimization problem in Eq. (10) involves only side constraints on the design variables, i.e., $\boldsymbol{x} \in \boldsymbol{X}$. Based on these

features, consider the sequence of intermediate distributions

$$\begin{aligned} \overline{p}_0(\boldsymbol{x}) &= U_{\boldsymbol{X}}(\boldsymbol{x}) \\ \overline{p}_j(\boldsymbol{x}) &\propto U_{\boldsymbol{X}}(\boldsymbol{x}) \exp\left(-\frac{h(\boldsymbol{x})}{T_j}\right),\ j = 1, 2, \ldots \end{aligned} \tag{12}$$

where $U_{\boldsymbol{X}}(\boldsymbol{x})$ represents a uniform distribution over the set $\boldsymbol{X}$. In this case, samples at the initial stage ($j = 0$) can be generated directly, while samples at the final stage ($T_j \to 0$) approximately follow a uniform distribution over the feasible set $\overline{\boldsymbol{X}}$. The TMCMC method [25] is implemented to generate the required samples. To achieve a smooth transition between distributions, the temperature parameter T_{j+1} satisfies the condition [16, 28]

$$\frac{\sum_{\ell=1}^{n} \exp\left(-h\left(\boldsymbol{x}_j^{(\ell)}\right)\left(T_{j+1}^{-1} - T_j^{-1}\right)\right)}{\left[\sum_{\ell=1}^{n} \exp\left(-h\left(\boldsymbol{x}_j^{(\ell)}\right)\left(T_{j+1}^{-1} - T_j^{-1}\right)\right)\right]^2} = \frac{1}{\nu n} \tag{13}$$

where $\boldsymbol{x}_j^{(\ell)}$, $\ell = 1, \ldots, n$, are n samples following $\overline{p}_j(\boldsymbol{x})$ and $\nu \in (0, 1)$ is a user-defined parameter. It is noted that all feasible designs generated during the intermediate stages of the sampling process are uniformly distributed over $\overline{\boldsymbol{X}}$ [14]. Therefore, the sampling process is stopped when $n_{feasible} \geq n_{target}$, where $n_{feasible}$ is the total number of feasible designs obtained during the different stages and n_{target} is a user-defined target value. At the end of the exploration phase, a total of $n_{feasible}$ designs uniformly distributed over $\overline{\boldsymbol{X}}$ are available.

3.4 Exploitation Phase

Starting from the set of feasible designs obtained during the exploration phase, which are distributed according to $p_0(\boldsymbol{x}) = U_{\overline{\boldsymbol{X}}}(\boldsymbol{x})$, the exploitation phase ultimately generates a set of designs lying in the vicinity of the optimal solution set $\boldsymbol{X}_f^*$. In this setting, samples following the intermediate distributions $p_j(\boldsymbol{x})$, $j = 1, 2, \ldots$, in Eq. (9) are obtained using the TMCMC method. The temperature parameter T_{j+1} verifies the relationship [16]

$$\frac{\sum_{\ell=1}^{n} \exp\left(-f\left(\boldsymbol{x}_j^{(\ell)}\right)\left(T_{j+1}^{-1} - T_j^{-1}\right)\right)}{\left[\sum_{\ell=1}^{n} \exp\left(-f\left(\boldsymbol{x}_j^{(\ell)}\right)\left(T_{j+1}^{-1} - T_j^{-1}\right)\right)\right]^2} = \frac{1}{\nu n} \tag{14}$$

where $\boldsymbol{x}_j^{(\ell)}$, $\ell = 1, \ldots, n$, are n samples following the distribution $p_j(\boldsymbol{x})$, and ν has been previously defined. As already pointed out, the distribution becomes uniform over the optimal solution set when $T_j \to 0$. For numerical implementation, however, a suitable stopping rule must be imposed. In this regard, the optimization procedure

is finished if (i) a prescribed maximum number of stages, N_{max}, are completed, or (ii) the sample coefficient of variation (c.o.v.) of the objective function is sufficiently small. Specifically, the sampling process stops at stage $j = 0, 1, \ldots$, if $j+1 = N_{max}$ or, alternatively, $\delta_{j+1} < \gamma\delta_0$, where $\gamma \in (0, 1)$ is a user-defined parameter and

$$\delta_j = \sqrt{\frac{1}{n-1}\sum_{\ell=1}^{n}\left(f\left(\boldsymbol{x}_j^{(\ell)}\right) - \left[\frac{1}{n}\sum_{i=1}^{n} f\left(\boldsymbol{x}_j^{(i)}\right)\right]\right)^2} \Bigg/ \left(\frac{1}{n}\sum_{\ell=1}^{n} f\left(\boldsymbol{x}_j^{(\ell)}\right)\right) \quad (15)$$

is the sample c.o.v. of the objective function $f(\boldsymbol{x})$ during stage j. The previous conditions indicate that the algorithm runs until a prescribed number of stages are completed or until δ_{j+1} becomes smaller than some fraction of the initial sample c.o.v. of the objective function, δ_0. It is noted that alternative stopping criteria can be implemented as well.

The samples $\left\{\boldsymbol{x}_{j+1}^{(1)}, \ldots, \boldsymbol{x}_{j+1}^{(n)}\right\}$ obtained at the final stage of the procedure can be regarded as a set of close-to-optimal designs. Thus, the proposed approach provides, in general, designs which are similar between each other in terms of their objective function values. This is particularly useful, e.g., in cases with multiple sub-optimal regions. Nevertheless, if a single solution is needed, the sample with the smallest objective function value can be selected.

3.5 Remarks

According to the previously described procedure, a set of close-to-optimal designs are obtained in a two-phase sampling framework in which the feasible and optimal solution sets are sequentially explored. In this regard, the proposed approach presents several advantageous features. First, due to its theoretical foundations and annealing properties, the procedure has high chances of reaching a vicinity of the optimal solution set. This includes cases involving multiple local optima, multiple discontinuous sub-feasible regions, and complex feasible design spaces [16]. Second, the formulation of the approach does not impose restrictions on the number of constraints or the behavior of the objective and constraint functions. Moreover, the proposed approach is not limited to a particular type of reliability assessment techniques. Thus, the method is quite general in the sense that, in principle, different classes of RBO problems can be treated with the same formulation. Third, the approach is suitable for practical implementation. In this regard, the same basic framework (i.e., the TMCMC method) is used in the exploration and exploitation phases. In addition, few user-defined parameters are required and, since the treatment of the reliability and standard constraints is direct, no special constraint-handling techniques such as penalty factors are necessary. Fourth, the method produces a set of nearly optimal designs rather than a single final solution. This can introduce additional flexibility to decision-making processes since an appropriate final design can be selected from

this set based on alternative considerations. Finally, the designs produced during the different stages of the sampling process allow to assess the sensitivity of the problem functions with respect to the design variables. This information, which is a byproduct of the procedure, gives a valuable insight into the RBO problem at hand.

4 Implementation Aspects

4.1 Transitional Markov Chain Monte Carlo Method

The TMCMC method [25], which has proved effective in several model updating applications (see, e.g., [30–33]), is implemented to carry out the exploration and exploitation phases. This technique draws samples at stage $j+1$, i.e., $\left\{\boldsymbol{x}_{j+1}^{(1)}, \ldots, \boldsymbol{x}_{j+1}^{(n)}\right\}$, by generating several Markov chains with stationary distribution equal to $\overline{p}_{j+1}$ (exploration phase) or p_{j+1} (exploitation phase). To this end, importance sampling concepts and the Metropolis-Hastings (M-H) algorithm are integrated [25]. In this setting, the ith sample, i.e., $\boldsymbol{x}_{j+1}^{(i)}$, is generated as follows:

- **Step 1:** Select a lead sample, $\widetilde{\boldsymbol{x}}$, as a design from the previous stage, $\boldsymbol{x}_j^{(\ell)}$, drawn with probability equal to its normalized importance weight $\overline{w}_j^{(\ell)} = w_j^{(\ell)} / \sum_{k=1}^{n} w_j^{(k)}$, $\ell = 1, \ldots, n$. The weights are given by $w_j^{(\ell)} = \overline{p}_{j+1}\left(\boldsymbol{x}_j^{(\ell)}\right) / \overline{p}_j\left(\boldsymbol{x}_j^{(\ell)}\right)$ for the exploration phase and $w_j^{(\ell)} = p_{j+1}\left(\boldsymbol{x}_j^{(\ell)}\right) / p_j\left(\boldsymbol{x}_j^{(\ell)}\right)$ for the exploitation phase. If the selected sample, $\boldsymbol{x}_j^{(\ell)}$, has been previously drawn, then the last state of its corresponding chain is selected as the lead sample $\widetilde{\boldsymbol{x}}$.
- **Step 2:** Draw a candidate design $\boldsymbol{x}^\star$ from the proposal distribution $p^*\left(\boldsymbol{x} | \widetilde{\boldsymbol{x}}\right)$. An adaptive proposal distribution with independent continuous and discrete components is considered [15, 16].
- **Step 3:** Set $\boldsymbol{x}_{j+1}^{(i)} = \boldsymbol{x}^\star$ with probability $\rho^* = \min\{1, \alpha\}$, where $\alpha = \overline{p}_{j+1}(\boldsymbol{x}^\star) p^*\left(\widetilde{\boldsymbol{x}} | \boldsymbol{x}^\star\right) / \overline{p}_{j+1}\left(\widetilde{\boldsymbol{x}}\right) p^*\left(\boldsymbol{x}^\star | \widetilde{\boldsymbol{x}}\right)$ for the exploration phase and $\alpha = p_{j+1}(\boldsymbol{x}^\star) p^*\left(\widetilde{\boldsymbol{x}} | \boldsymbol{x}^\star\right) / p_{j+1}\left(\widetilde{\boldsymbol{x}}\right) p^*\left(\boldsymbol{x}^\star | \widetilde{\boldsymbol{x}}\right)$ for the exploitation phase. If the candidate state is rejected, the lead sample is repeated, i.e., set $\boldsymbol{x}_{j+1}^{(i)} = \widetilde{\boldsymbol{x}}$.

The previous procedure is iteratively carried out until the required number of samples has been obtained. A more detailed description of the TMCMC method, from the theoretical and implementation viewpoints, can be found in [25].

4.2 Proposal Distribution

The proposed approach requires, in the context of the M-H algorithm, the definition of an appropriate proposal distribution to draw samples from the intermediate distributions. As already pointed out, an adaptive proposal distribution with independent continuous and discrete components is considered in this work [15, 16]. That is, the candidate state is generated from a distribution $p^*\left(\boldsymbol{x}|\,\widetilde{\boldsymbol{x}}\right) = p_c^*\left(\boldsymbol{x}_c|\widetilde{\boldsymbol{x}}_c\right) p_d^*\left(\boldsymbol{x}_d|\widetilde{\boldsymbol{x}}_d\right)$, where $p_c^*\left(\boldsymbol{x}_c|\widetilde{\boldsymbol{x}}_c\right)$ is the proposal distribution for the continuous design variables and $p_d^*\left(\boldsymbol{x}_d|\widetilde{\boldsymbol{x}}_d\right)$ for the discrete design variables.

4.2.1 Continuous Design Variables

Following some of the ideas in [25], the proposal distribution for the continuous variables, $p_c^*\left(\boldsymbol{x}_c|\widetilde{\boldsymbol{x}}_c\right)$, is a Gaussian distribution centered at the continuous components of the lead sample. The corresponding covariance matrix, $\widetilde{\boldsymbol{\Sigma}}$, is taken as

$$\widetilde{\boldsymbol{\Sigma}} = \beta^2 \sum_{\ell=1}^{n} \overline{w}_j^{(\ell)} \left(\boldsymbol{x}_{j,c}^{(\ell)} - \overline{\boldsymbol{x}}_{j,c}\right)\left(\boldsymbol{x}_{j,c}^{(\ell)} - \overline{\boldsymbol{x}}_{j,c}\right)^T \tag{16}$$

where β is a scaling parameter, $\boldsymbol{x}_{j,c}^{(\ell)}$, $\ell = 1, \ldots, n$, are the continuous components of the samples at stage j, $\overline{\boldsymbol{x}}_{j,c} = \sum_{\ell=1}^{n} \overline{w}_j^{(\ell)} \boldsymbol{x}_{j,c}^{(\ell)}$, and the normalized weights $\overline{w}_j^{(\ell)}$, $\ell = 1, \ldots, n$, have been previously defined. The scaling parameter β is adaptively tuned according to the observed acceptance rate of the M-H algorithm [34].

4.2.2 Discrete Design Variables

The discrete proposal distribution, $p_d^*\left(\boldsymbol{x}_d|\widetilde{\boldsymbol{x}}_d\right)$, considers the discrete components to be independent between each other, i.e.,

$$p_d^*\left(\boldsymbol{x}_d|\widetilde{\boldsymbol{x}}_d\right) = \prod_{i=1}^{n_d} p_{di}^*(x_{di}|\tilde{x}_{di}) \tag{17}$$

where $\widetilde{\boldsymbol{x}}_d = \langle \tilde{x}_{d1}, \ldots, \tilde{x}_{dn_d} \rangle^T$ contains the discrete components of the lead sample and $p_{di}^*(x_{di}|\tilde{x}_{di})$ is the proposal distribution corresponding to the ith discrete component. The definition of this distribution relies on the set of neighbors of the current lead sample, $\tilde{x}_{di}$, within the corresponding set of available values $\boldsymbol{X}_{di}$. This set is defined as

$$\mathrm{Adj}(\tilde{x}_{di}) = \left\{ x_{di(m)}, m = 1, \ldots, n_{di} : \lambda\left(\tilde{x}_{di}, x_{di(m)}\right) \leq \lambda_i^* \right\} \tag{18}$$

where $\lambda\left(\tilde{x}_{di}, x_{di(m)}\right)$ is the *distance* between $\tilde{x}_{di}$ and $x_{di(m)}$ within the set $\boldsymbol{X}_{di}$, and λ_i^* is a user-defined parameter. This distance is defined in terms of the indices of the values within the sorted set $\boldsymbol{X}_{di}$. For instance, if $\tilde{x}_{di}$ is equal to the sth available value, i.e., $\tilde{x}_{di} = x_{di(s)}$, then the distance measure becomes $\lambda\left(\tilde{x}_{di}, x_{di(m)}\right) = \lambda\left(x_{di(s)}, x_{di(m)}\right) = |s - m|$. Based on the previous definitions, the proposal distribution for the ith component is given by [15, 16]

$$p_{di}^*(x_{di}|\tilde{x}_{di}) = \begin{cases} \frac{1-\tau_i}{\mathrm{Card}\left[\mathrm{Adj}(\tilde{x}_{di})\right]}, & \text{if } x_{di} \in \mathrm{Adj}(\tilde{x}_{di}) \\ \frac{\tau_i}{n_{di} - \mathrm{Card}\left[\mathrm{Adj}(\tilde{x}_{di})\right]}, & \text{otherwise} \end{cases} \tag{19}$$

where $\tau_i \in [0, 1]$ represents the probability of randomly selecting a discrete value that does not belong to the set of neighbors of the lead sample $\tilde{x}_{di}$, and Card$[\cdot]$ is the cardinality of the set within square brackets.

In the adopted proposal distribution, the parameters τ_i and λ_i^* jointly characterize its corresponding spread. Such parameters can be directly defined by the user or adaptively modified during the sampling process. In particular, an adaptive strategy is implemented here to update the value of λ_i^* at the beginning of each stage of the exploitation phase. First, the *maximum number of consecutive elements* of $\boldsymbol{X}_{di}$ that were observed during the previous stage, denoted by η, is identified. Then, the parameter λ_i^* is defined as

$$\lambda_i^* \leftarrow \min\left\{\lambda_i^*, L\right\} \tag{20}$$

with L the largest integer such that $L \leq (\eta - 1)/2$. The updating rule is repeated for $i = 1, \ldots, n_d$. This strategy tends to reduce the value of λ_i^* for advanced exploitation stages, which can improve the sampling efficiency. Nonetheless, alternative updating approaches can also be considered.

4.2.3 Acceptance Probability

As previously mentioned, the candidate state $\boldsymbol{x}^\star = \langle \boldsymbol{x}_c^{\star T}, \boldsymbol{x}_d^{\star T} \rangle^T$, which is drawn from $p^*\left(\boldsymbol{x}|\,\widetilde{\boldsymbol{x}}\right)$, becomes the next state of the Markov chain with probability $\rho^* = \min\{1, \alpha\}$. For the exploration phase, the quantity α is given by

$$\alpha = I\left[\boldsymbol{x}^\star \in \boldsymbol{X}\right] \frac{\exp\left(-\frac{h(\boldsymbol{x}^\star)}{T_{j+1}}\right)}{\exp\left(-\frac{h\left(\widetilde{\boldsymbol{x}}\right)}{T_{j+1}}\right)} \frac{p_d^*\left(\widetilde{\boldsymbol{x}}_d|\boldsymbol{x}_d^\star\right)}{p_d^*\left(\boldsymbol{x}_d^\star|\widetilde{\boldsymbol{x}}_d\right)} \tag{21}$$

where $I[\cdot] = 1$ if the expression within square brackets is true and $I[\cdot] = 0$ otherwise. In addition, for the exploration phase the quantity α becomes

$$\alpha = I\left[\boldsymbol{x}^\star \in \overline{\boldsymbol{X}}\right] \frac{\exp\left(-\frac{f(\boldsymbol{x}^\star)}{T_{j+1}}\right)}{\exp\left(-\frac{f\left(\widetilde{\boldsymbol{x}}\right)}{T_{j+1}}\right)} \frac{p_d^*\left(\widetilde{\boldsymbol{x}}_d|\boldsymbol{x}_d^\star\right)}{p_d^*\left(\boldsymbol{x}_d^\star|\widetilde{\boldsymbol{x}}_d\right)} \tag{22}$$

In the previous expressions, the ratio $p_d^*\left(\widetilde{\boldsymbol{x}}_d|\boldsymbol{x}_d^\star\right)/p_d^*\left(\boldsymbol{x}_d^\star|\widetilde{\boldsymbol{x}}_d\right)$ is given by

$$\frac{p_d^*\left(\widetilde{\boldsymbol{x}}_d|\boldsymbol{x}_d^\star\right)}{p_d^*\left(\boldsymbol{x}_d^\star|\widetilde{\boldsymbol{x}}_d\right)} = \prod_{i=1}^{n_d} \begin{cases} \frac{\mathrm{Card}\left[\mathrm{Adj}(\tilde{x}_{di})\right]}{\mathrm{Card}\left[\mathrm{Adj}(x_{di}^\star)\right]}, & \text{if } x_{di}^\star \in \mathrm{Adj}(\tilde{x}_{di}) \\ \frac{n_{di} - \mathrm{Card}\left[\mathrm{Adj}(\tilde{x}_{di})\right]}{n_{di} - \mathrm{Card}\left[\mathrm{Adj}(x_{di}^\star)\right]}, & \text{otherwise} \end{cases} \tag{23}$$

4.3 Adaptive Surrogate Model

The proposed approach requires the sequential generation of samples. Consequently, a significant number of first-passage probability evaluations may be required by the optimization procedure. To alleviate the corresponding numerical demands, an adaptive surrogate model based on kriging interpolants is implemented [17, 35]. These metamodels approximate the target function using an underlying Gaussian process whose properties depend on the available data points [36, 37]. Some of their advantages are that they do not require a regular grid of support points, the c.o.v. of the kriging prediction can be directly estimated, and they are exact at the support points. Furthermore, given the annealing nature of the proposed approach, the effective support of the current distribution is generally contained in that of prior stages. Thus, data points from previous stages can be used to construct the metamodel at the current stage. As a result, the previous features enable a local and adaptive surrogate model in which (i) the support points that lie closer to the candidate design are used, and (ii) the database is enriched as new designs in the important region of the target distribution are generated.

Specifically, consider a first-passage probability function, $P_F(\boldsymbol{x})$, involved in the definition of one of the problem functions. A kriging metamodel, $\widehat{p}^{\mathrm{kr}}(\boldsymbol{x})$, is implemented to approximate the logarithm of this function, i.e., $\widehat{p}_F^{\mathrm{kr}}(\boldsymbol{x}) \approx \ln P_F(\boldsymbol{x})$. At the beginning of the sampling process, a database corresponding to full model evaluations of $\ln P_F(\boldsymbol{x})$ is obtained by means of any suitable strategy. Then, the surrogate prediction of a candidate design $\boldsymbol{x}^\star$ corresponding to a given Markov chain is obtained as follows.

- **Step 1:** Find the N_{sp} points in the database of available values of $\ln P_F(\boldsymbol{x})$ that are closer to the starting seed of the current Markov chain. The Euclidian distance

is considered to this end, although alternative distance metrics can be adopted as well. Construct the kriging metamodel for the current Markov chain, $\widehat{p}_F^{\mathrm{kr}}(\boldsymbol{x})$, using this set of designs as the corresponding support points.

- **Step 2:** For a given candidate design, $\boldsymbol{x}^\star$, verify the following criteria:
  - **Step 2-a:** Check if $\boldsymbol{x}^\star$ belongs to the n_x-dimensional convex hull of the support points. If not, go to step 4.
  - **Step 2-b:** Compute the kriging prediction $\widehat{p}^{\mathrm{kr}}(\boldsymbol{x}^\star)$. If this value is smaller than the Q-quantile of $\ln P_F(\boldsymbol{x})$ in the database, go to step 4.
  - **Step 2-c:** Compute the c.o.v. of the kriging estimate. If this value is larger than a user-defined tolerance $\epsilon > 0$, go to step 4.
- **Step 3:** If all the criteria in step 2 are verified, the kriging prediction is accepted. Set $P_F(\boldsymbol{x}) = \exp\left(\widehat{p}_F^{\mathrm{kr}}(\boldsymbol{x})\right)$ and continue the sampling process.
- **Step 4:** If at least one criterion in step 2 does not hold, the kriging prediction is rejected. An exact evaluation of $P_F(\boldsymbol{x}^\star)$ is performed and this point is added to the database. Continue the sampling process.

The criteria in step 2 aim to control the quality of the kriging estimate. In addition, the set of support points is kept fixed throughout the generation of a given Markov chain. This is done to maintain the Markovian property of the chain and to avoid potential discontinuities associated with slightly different sets of support points. A more detailed description of the strategy can be found in [13, 17, 35].

4.4 Parallelization Features

High-performance computing techniques at the computer hardware level can be considered for improved computational efficiency. In this regard, the proposed approach is particularly suitable for parallel implementation due to the properties of the TMCMC method. First, the initial stage of the exploration phase corresponds to direct Monte Carlo simulation and, therefore, it can be fully scheduled in parallel. Thereafter, the method produces Markov chains that can be generated independently. Since the numerical cost of evaluating the reliability at each design is difficult to predict, dynamic scheduling schemes can be beneficial to distribute the function evaluations on a first-come-first-serve basis [17, 35].

Parallelization strategies can also be integrated with the use of adaptive surrogate models to enhance the numerical efficiency of the proposed approach. To this end, a total of n_{par} samples are generated simultaneously, and then the database of support points is updated. This procedure is repeated during each stage until the required sample size is reached. This allows exploiting the parallel features of the TMCMC method while enriching the kriging database on a regular basis. The parameter n_{par} should be relatively small to promote the adaptability of the surrogate model, but not smaller than a certain value beyond which it becomes detrimental to the efficaciousness of the parallelization process.

5 Application Examples

Three application examples involving structural dynamical systems under stochastic excitation are presented, which include unconstrained, constrained, discrete and continuous formulations. In all cases, the structural behavior can be modeled by a multi-degree of freedom system satisfying the equation of motion

$$\boldsymbol{M}\ddot{\boldsymbol{y}}(t) + \boldsymbol{C}\dot{\boldsymbol{y}}(t) + \boldsymbol{K}\boldsymbol{y}(t) + \boldsymbol{\kappa}_{NL}(\boldsymbol{y}(t), \dot{\boldsymbol{y}}(t), \boldsymbol{q}(t)) = \boldsymbol{f}(t) \tag{24}$$

where $\boldsymbol{y}(t)$ denotes the displacement vector, $\boldsymbol{\kappa}_{NL}(\boldsymbol{y}(t), \dot{\boldsymbol{y}}(t), \boldsymbol{q}(t))$ is the vector of nonlinear restoring forces, $\boldsymbol{q}(t)$ comprises the state variables of the nonlinear components, and $\boldsymbol{f}(t)$ represents the excitation vector. The matrices $\boldsymbol{M}$, $\boldsymbol{C}$, and $\boldsymbol{K}$ characterize, respectively, the mass, damping, and stiffness of the system. In addition, the evolution of $\boldsymbol{q}(t)$ depends on an appropriate nonlinear model. Thus, finding the system response requires, in general, to solve a set of coupled nonlinear equations using suitable time integration schemes. Subset simulation [38, 39] is implemented for reliability assessment in all examples. Nonetheless, alternative simulation methods can also be considered.

5.1 Example 1

The reliability-based design of a bridge system subject to stochastic ground excitation is addressed. Figure 1 presents an isometric view of the structural model, which has been borrowed from [13, 40]. The bridge is curved in plan and has five spans of lengths equal to 27, 25, 23, 20, and 24 m, which give a total length of 119 m. The deck is monolithically supported by four piers of 8 m height, and each pier is founded on an array of four piles of 35 m height. To model the soil-pile interaction, a series of linear translational springs are incorporated along the height of each pile, with stiffness constants increasing linearly from 560 T/m at the surface to 11,200 T/m at the base. The piers and piles are modeled as column elements with diameters of 1.6 and 0.6 m, respectively. In addition, the deck cross section is a box girder modeled with beam and shell elements. Two sliding bearings are included at each abutment to support the bridge deck. These nonlinear devices consist of an upper steel plate with a housing cap for the slider, a bottom plate with a concave semi-spherical stainless-steel surface, and a steel slider [41, 42]. A sketch of the sliding bearing is also shown in Fig. 1. Overall, the finite element model of the bridge involves more than 10,000 degrees of freedom.

For dynamical analysis purposes, it is assumed that the nonlinearities are localized in the response of the sliding bearings, while the piers, piles, and deck remain linear. In this regard, the linear components are characterized by an elastic modulus equal to 2.94×10^{10} N/m^2, a Poisson ratio equal to 0.2, and a density of 2500 kg/m^3. In addition, a 5% of critical damping ratio is considered. On the other hand, the sliding

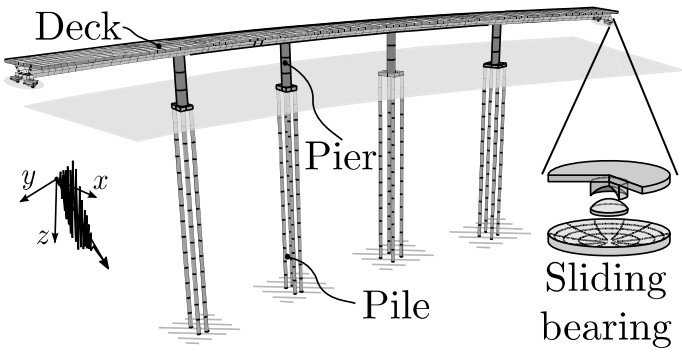


Fig. 1 Isometric view of the nonlinear bridge model. Example 1

bearings at the abutments are characterized by an experimentally validated model that incorporates performance degradation effects. These relate to changes in the friction coefficient due to variations in the vertical load, in the relative velocity between plates, and in the sliding surface temperature [42]. For illustration purposes, a representative displacement-restoring force curve of these devices is shown in Fig. 2.

As shown in Fig. 1, the structural model is subject to a seismic excitation applied at 40° with respect to the y axis. A point-source model [43, 44] is considered to characterize the ground excitation as a non-stationary stochastic process. This class of models links available knowledge about the geological site with the uncertainty of future ground motions. To this end, a white noise sequence is considered along with several seismic parameters [43–46]. The duration of the excitation is $t_T = 10$ s with a time step of 0.01 s. Thus, more than 1000 random variables are involved in the characterization of the excitation.

In this example, the goal is to minimize the failure probability of the system subject to side constraints on the design variables [13]. Formally, the corresponding RBO problem is formulated as

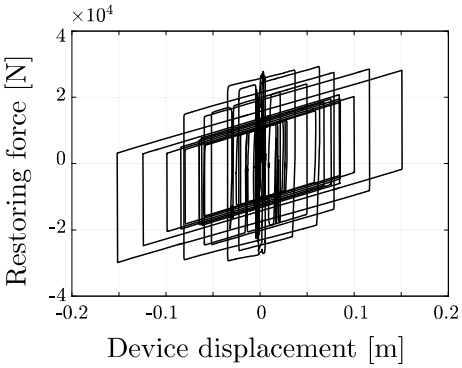


Fig. 2 Typical displacement-restoring force curve of the devices. Example 1

$$\begin{aligned} &\min_{\boldsymbol{x}=\langle x_1,x_2\rangle^T} \quad P_F(\boldsymbol{x}) \\ &\text{s.t.} \qquad\quad 0.5 \leq x_i \leq 1.5,\ i=1,2 \end{aligned} \tag{25}$$

where x_1 and x_2 are the design variables, and $P_F(\boldsymbol{x})$ is the probability of $F = \{d(\boldsymbol{x},\boldsymbol{\theta}) > 1\}$ evaluated at $\boldsymbol{x}$. The corresponding demand function is

$$d(\boldsymbol{x},\boldsymbol{\theta}) = \max_{\ell=1,2,3} \max_{t\in[0,T]} \frac{|a_\ell(t,\boldsymbol{x},\boldsymbol{\theta})|}{a_\ell^*} \tag{26}$$

where $a_\ell(t,\boldsymbol{x},\boldsymbol{\theta})$ is the absolute acceleration at the ℓth control point and $a_\ell^* = 6.9\ \text{m/s}^2$. The control points are located at the two abutments and the deck midpoint.

The design variables are associated with two model parameters of the sliding bearings, namely, the initial friction coefficient (μ_0) and the radius of the concave surface (R). For optimization purposes, the intermediate design variables are specified as $x_1 = \mu_0/\overline{\mu}_0$ and $x_2 = R/\overline{R}$ with reference values $\overline{\mu}_0 = 0.106$ and $\overline{R} = 2.235$ m. The initial friction coefficient affects the dissipation capacity of the bearing system, while the radius of the concave surface controls its natural frequency. Thus, it is expected that both parameters will significantly affect the system behavior.

To illustrate the effect of the design variables on the objective function, Fig. 3 presents contours of $P_F(\boldsymbol{x})$ in the design space. The results indicate that the failure probability tends to decrease for higher values of the initial friction coefficient and lower values of the radius of the concave surface. Further, a close inspection of the plot shows that the optimal solution set seems to involve a valley near the lower-right corner of the design space. For clarity, this region is highlighted in the figure. Thus, multiple solutions with very similar objective function values can be expected in this case.

As previously pointed out, the RBO problem in Eq. (25) involves only side constraints on the design variables. Hence, uniformly distributed feasible designs can be obtained directly and only the exploitation phase is needed in this case. The proposed approach is implemented with $n = 500$ samples per stage and $\nu = 0.5$.

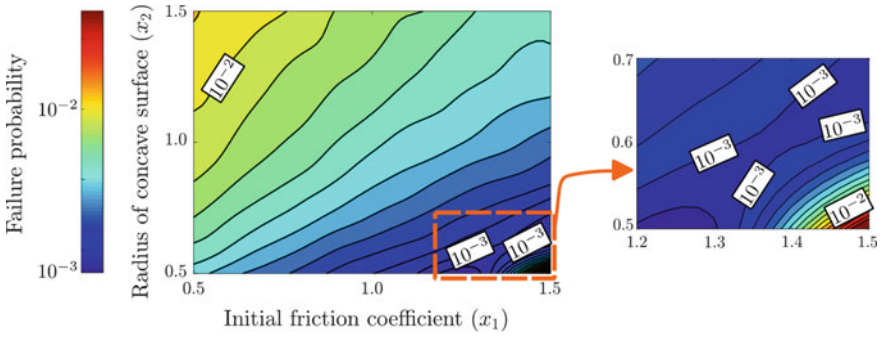


Fig. 3 Iso-probability curves in the design space. Example 1

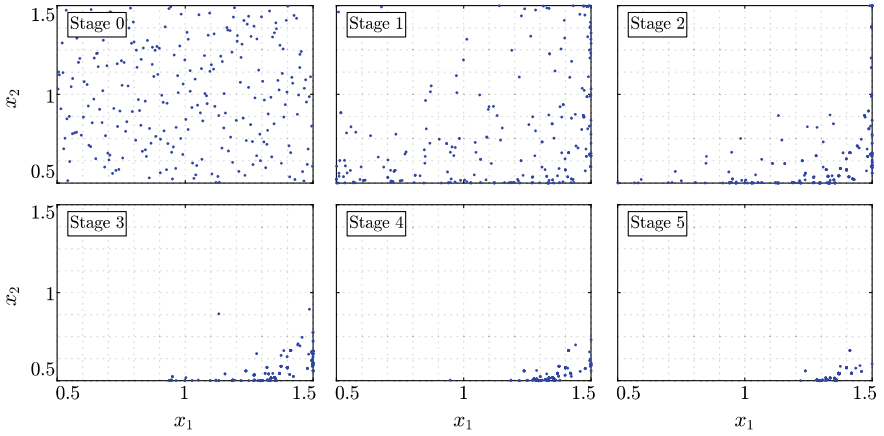


Fig. 4 Designs obtained during the different optimization stages. Example 1

In addition, the adaptive surrogate model is implemented with 20 support points, $\epsilon = 0.1$, and $Q = 0.05$. The database is initialized with all the designs from the initial stage. However, alternative initialization strategies can also be considered.

Figure 4 presents the samples obtained at the different stages of the exploitation phase. The initial designs (stage 0) are uniformly distributed over the design space, whereas the samples at the final stage lie in a region near the lower-right corner of the design space. The final set of designs resembles the optimal region identified in Fig. 3, which illustrates the effectiveness of the approach. The corresponding failure probability values roughly range between 3×10^{-4} and 7×10^{-4}. The sample-based optimal design is given by $\boldsymbol{x}^T = \langle 1.32, 0.51 \rangle$ with $P_F(\boldsymbol{x}) = 3.17 \times 10^{-4}$.

As previously pointed out, an adaptive surrogate model is integrated for improved numerical efficiency. To assess its performance, Fig. 5 shows the acceptance rate of the kriging predictions after the initial stage. This quantity corresponds to the fraction of designs that satisfy all acceptance criteria, and it remains above 80% throughout the optimization procedure. Overall, after the initial stage (stage 0), no more than 8% of the total number of design evaluations are performed with full reliability assessment. Thus, the use of kriging greatly improves the numerical efficiency of the method without sacrificing the quality of the optimization results.

5.2 Example 2

The design of a nonlinear 52-story reinforced concrete building is considered as the second example. The structural model, which has been borrowed from [14], comprises more than 50,000 degrees of freedom. For reference purposes, an isometric view of the building and the plan view of each floor are presented in Fig. 6. The building includes a core of shear walls and a perimeter of columns with circular cross

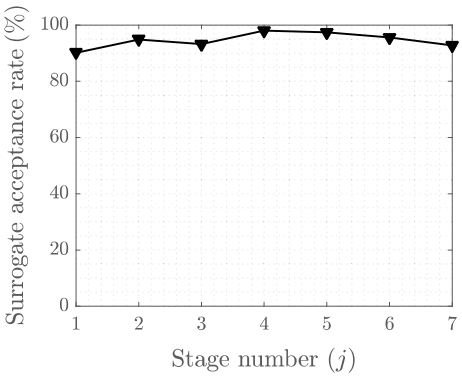


Fig. 5 Metamodel acceptance rate during the optimization stages. Example 1

sections. The nominal value for the corresponding thicknesses and diameters is equal to 0.40 m, whereas the slab thickness is equal to 0.20 m. The same material properties from the previous example are considered. For improved seismic performance, four nonlinear hysteretic devices are placed at each floor to increase the stiffness and dissipation capacity of the system. As indicated in Fig. 6, these devices are located along the axes 4, 7, 8 and 11. The restoring force in each device is given by $\kappa(t) = k^e v(t)$ where $k^e = 2.8 \times 10^9$ N/m is the initial stiffness, $v(t) = \delta(t) - q_1(t) + q_2(t)$, $\delta(t)$ denotes the interstory displacement, and $q_1(t)$ and $q_2(t)$ are the plastic elongations of the device. These variables satisfy the first-order nonlinear differential equations [47]

$$\begin{aligned}
\dot{q}_1(t) &= \dot{\delta}(t) H\big(\dot{\delta}(t)\big)\left[H\big(v(t) - v_y\big) \frac{v(t) - v_y}{v_p - v_y} H\big(v_p - v(t)\big) + H\big(v(t) - v_p\big) \right] \\
\dot{q}_2(t) &= -\dot{\delta}(t) H\big(-\dot{\delta}(t)\big) \\
&\quad \left[H\big(-v(t) - v_y\big) \frac{-v(t) - v_y}{v_p - v_y} H\big(v_p + v(t)\big) + H\big(-v(t) - v_p\big) \right]
\end{aligned} \tag{27}$$

where $H(\cdot)$ is the Heaviside step function, $v_y = 0.0042$ m is the yielding onset, and $\kappa_p = k^e v_p$ is the maximum restoring force of the device with $v_p = 0.006$ m.

As illustrated in Fig. 6 (right), the building is subject to a horizontal ground excitation, $\ddot{u}_g(t)$, acting along the y axis. This excitation is characterized as a stochastic process using a point-source model, as in the previous example. For dynamic analysis purposes, each floor is regarded as rigid within its plane when compared with the rest of structural components. Thus, by means of appropriate condensation techniques, the degrees of freedom of the entire model are linked to three coordinates per floor (one rotational and two translational displacements). In addition, a 5% of critical damping ratio at the modal level is considered.

In this case, the variables to be controlled correspond to the dimensions of the shear walls and exterior columns (see Fig. 6). In particular, the wall thickness (t_w) and column diameter (d_c) at each floor are linked to an intermediate optimization variable,

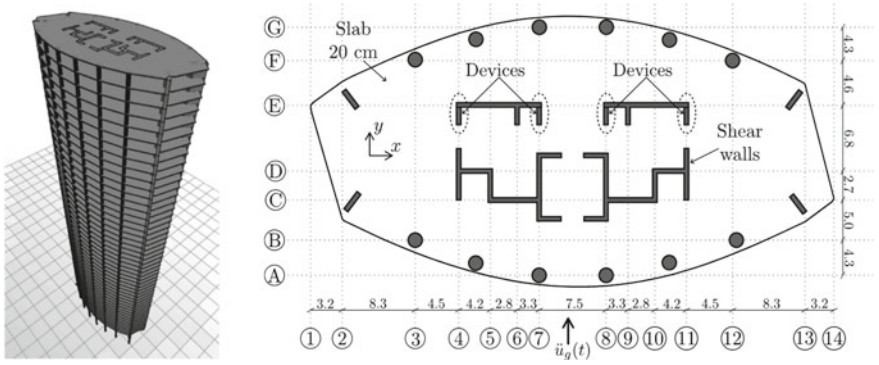


Fig. 6 Isometric view (left) and floor plan (right) of the 52-story building model. Example 2

x, as $t_w = \overline{t}_w x$ and $d_c = \overline{d}_c x$, respectively, with nominal values $\overline{t}_w = \overline{d}_c = 0.40$ m. For optimization purposes, a total of $n_x = 6$ intermediate variables are considered. These are linked to the design elements of different floors as follows: x_1 is associated with floors 1–9, x_2 with floors 10–18, x_3 with floors 19–26, x_4 with floors 27–35, x_5 with floors 36–44, and x_6 with floors 45–52. Then, a constrained optimization problem is formulated as

$$\begin{aligned} \min_{\boldsymbol{x}} \; & f(\boldsymbol{x}) = \sum_{i=1}^{6} x_i/6 \\ \text{s.t.} \; & P_{F_j}(\boldsymbol{x}) \le 10^{-3}, \; j = 1, 2 \\ & x_{i+1} \le x_i, \qquad i = 1, \ldots, 5 \\ & 0.5 \le x_i \le 1.5, \; i = 1, \ldots, 6 \end{aligned} \tag{28}$$

where the failure probability functions $P_{F_1}(\boldsymbol{x})$ and $P_{F_2}(\boldsymbol{x})$ are associated with the displacement at the first and top floors, respectively, and the standard constraints $x_{i+1} \le x_i$ impose that the dimensions of upper floor members cannot be larger than those of lower floors. The corresponding first-passage failure events are defined as $F_j = \left\{ d_j(\boldsymbol{x}, \boldsymbol{\theta}) > 1 \right\}$, $j = 1, 2$, where

$$d_j(\boldsymbol{x}, \boldsymbol{\theta}) = \max_{t \in [0,T]} \frac{\left| u_j^y(t, \boldsymbol{x}, \boldsymbol{\theta}) \right|}{u_j^*}, \quad j = 1, 2 \tag{29}$$

with $u_1^y(t, \boldsymbol{x}, \boldsymbol{\theta})$ and $u_2^y(t, \boldsymbol{x}, \boldsymbol{\theta})$ denoting the ground-relative displacement along the y direction at the centroid of the first and top floors, respectively. The corresponding thresholds u_1^* and u_2^* are equal to 0.08% of the first story height and 0.075% of the building height, respectively. The reference period is taken as $t_T = 15$ s with a time step of 0.01 s, which leads to more than 1500 random variables involved in the corresponding multidimensional probability integrals.

The proposed approach is implemented by considering $n = 500$ samples per stage and $\nu = 0.5$ for both phases. The exploration phase considers a target sample size of $n_{target} = 750$ as stopping criterion, whereas $N_{\max} = 10$ stages are carried out in the exploitation phase. In addition, the adaptive surrogate model strategy is implemented to approximate the two failure probability functions throughout the entire optimization process. In this regard, the corresponding kriging metamodel considers 28 support points, $\epsilon = 0.1$ and $Q = 0.05$.

First, an exploration phase is carried out. The corresponding final set of samples, which comprises 950 feasible designs, is shown in Fig. 7 in terms of two-dimensional projections and marginal histograms. These designs have been retrieved from the eight TMCMC stages carried out during the exploration phase and approximately follow a uniform distribution over the feasible set. The marginal histograms indicate that the effective support of the first design variable is smaller than that of, e.g., the sixth design variable. Thus, the dimensions of the core walls and columns of lower floors seem to be more relevant than of upper floors to determine the feasibility of a given design. Finally, validation calculations indicate that the hypervolume of the feasible set represents less than 0.01% of the initial search space. This illustrates the ability of the proposed approach to explore feasible design spaces with challenging geometries in an effective manner.

Starting from the set of samples in Fig. 7, an exploitation phase is carried out to explore the optimal solution set. The final designs, which are obtained after ten TMCMC stages, are shown in Fig. 8. It is seen that they are densely concentrated near a single value in the design space. In fact, the corresponding objective function values range between 6.333 and 6.357 at the final stage. These values represent a relative variation of less than 0.3% and, thus, they can be considered as equivalent from the optimization viewpoint. This illustrates one of the advantages of the proposed approach, as it yields a set of nearly optimal solutions which, in turn, provides additional flexibility for decision-making purposes. Nonetheless, if a single solution is needed, the design with the smallest objective function value can be chosen. Such design is $\boldsymbol{x}^T = \langle 1.489, 1.476, 1.088, 1.011, 0.751, 0.515 \rangle$, with $f(\boldsymbol{x}) = 6.333$, $P_{F_1}(\boldsymbol{x}) = 0.999 \times 10^{-3}$ and $P_{F_2}(\boldsymbol{x}) = 0.169 \times 10^{-3}$. In addition, the standard constraints verify $x_2/x_1 = 0.991$, $x_3/x_2 = 0.736$, $x_4/x_3 = 0.930$, $x_5/x_4 = 0.743$, and $x_6/x_5 = 0.685$. It is seen that the first reliability constraint, whose response of interest is the first story drift, and the first geometric constraint can be regarded as active at this design from a practical point of view.

As previously pointed out, an adaptive metamodel has been implemented. The initial database corresponds to the initial samples of the exploration phase. Numerical results indicate that the surrogate acceptance rate remains above 97% throughout the different stages of the exploration and exploitation phases, which yields an overall speedup factor of more than 16 in this case. Further, validation calculations show that the results obtained with and without the use of metamodels are very similar. Thus, the use of surrogate modeling techniques together with the proposed approach prove instrumental for the effective solution of a class of challenging RBO problems, such as those involving complex engineering systems.

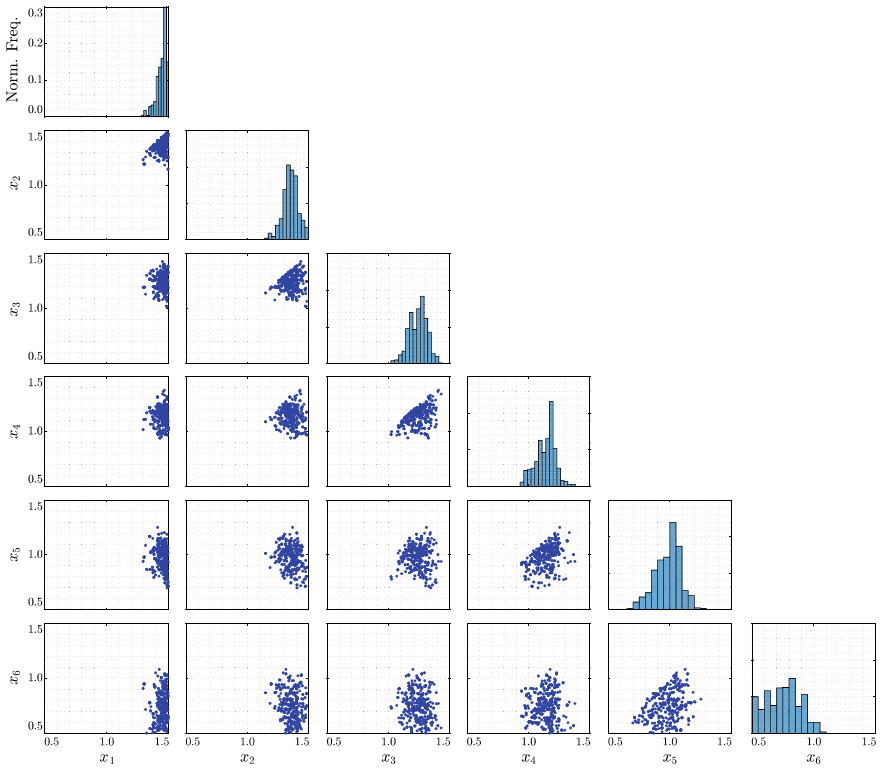


Fig. 7 Set of feasible designs obtained in the exploration phase. Example 2

5.3 Example 3

The design of the bracing system for a 4-story building is considered as the third example. The corresponding structural model, which has been borrowed from [15], is illustrated in Fig. 9. Each floor is supported by 48 identical columns. The corresponding cross sections, which are taken from AISC standards [48], are W24 × 131 for the two lower floors and W24 × 104 for the two upper floors. In addition, six nonlinear devices, which follow the same restoring force law from the previous example, are implemented at each floor. Finally, a bracing system consisting of 128 tubular steel elements is incorporated. The braces along axes A, C, D and F act in the x direction, while those along axes 1, 2, 7 and 8 in the y direction. The elastic modulus and density of the bracing elements are taken as 2.1×10^{11} N/m^2 and 7.42 ton/m^3, respectively.

The system is subject to a ground excitation applied at 45 degrees with respect to the x axis (see Fig. 9). As in the previous examples, the ground acceleration is represented using a point-source model. A reference period of $t_T = 15$ s is considered with a time step of 0.01 s and, therefore, more than 1500 random variables

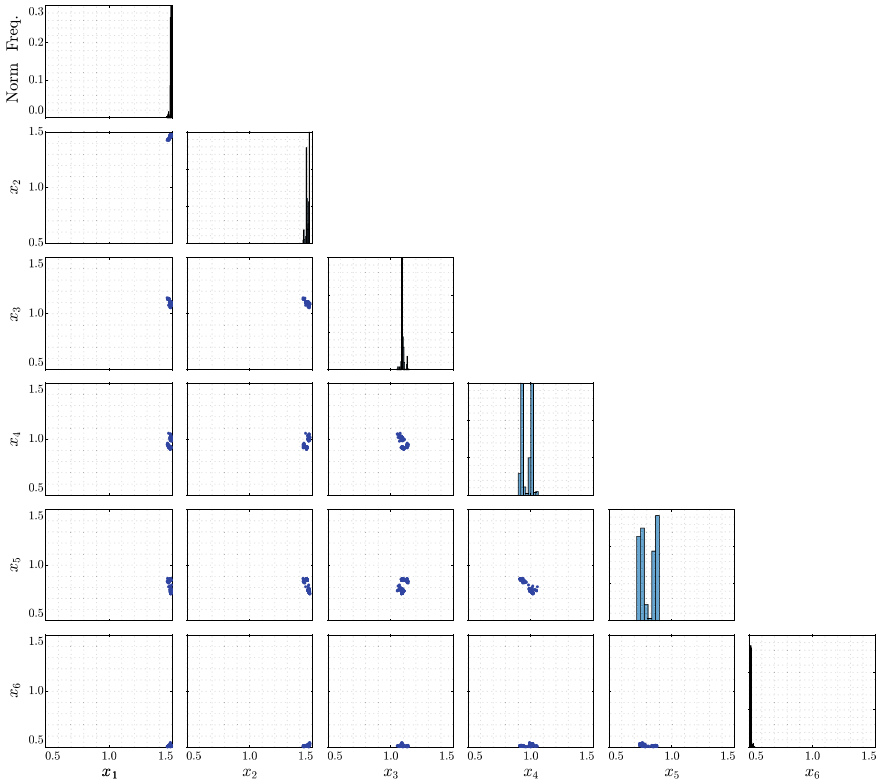


Fig. 8 Set of designs at the end of the exploitation phase. Example 2

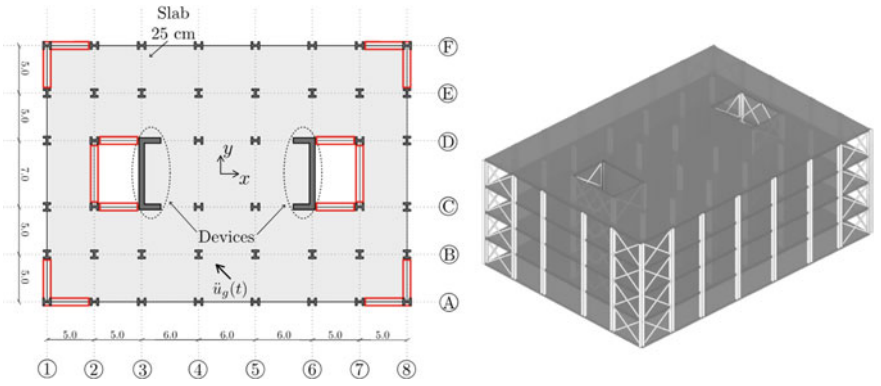


Fig. 9 Floor plan (left) and isometric view (right) of the 4-story building. Example 3

are involved in the characterization of the stochastic process. For dynamic analysis purposes, each floor is assumed as rigid within the $x - y$ plane when compared with the horizontal resistant elements. Using condensation techniques, the global system response can be characterized with three coordinates per floor. The mass and polar inertia of each floor are 5.98×10^5 kg and 1.10×10^8 kg m^2, respectively. Finally, a 2% of critical damping is considered in the model.

In this example, the objective function $f(\boldsymbol{x})$ is related to the total weight of the brace elements. A total of four design variables are considered to define the areas of the tubular cross sections of the bracing elements. The areas can be chosen from a discrete set of 48 available values ranging from 719 to 3063 mm^2 [15]. Each design variable x_i, $i = 1, \ldots, 4$, is linked to the brace elements of two consecutive floors along a certain direction. The variables x_1 and x_2 represent the areas of the bracing elements in the two lower floors along the x and y directions, respectively. For the two upper floors, the areas of the brace elements along the x and y directions are given by x_3 and x_4, respectively. In this setting, the initial search space comprises more than 5×10^6 available configurations for the bracing system. A constrained RBO problem is formulated as

$$\begin{aligned} \min_{\boldsymbol{x}} \; & f(\boldsymbol{x}) = \sum_{i=1}^{4} \overline{f}_i x_i \\ \text{s.t.} \; & P_{F_j}(\boldsymbol{x}) \leq 5 \times 10^{-4}, \; j = 1, 2 \\ & x_i \in \boldsymbol{X}, \qquad\qquad i = 1, \ldots, 4 \end{aligned} \tag{30}$$

where $f(\boldsymbol{x})$ represents the normalized weight of the bracing system with normalizing constants $\overline{f}_1 = \overline{f}_3 = 8.5 \times 10^{-5}$ and $\overline{f}_2 = \overline{f}_4 = 7.8 \times 10^{-5}$, $P_{F_j}(\boldsymbol{x})$ is the probability of failure event $F_j = \left\{ d_j(\boldsymbol{x}, \boldsymbol{\theta}) > 1 \right\}$, and $\boldsymbol{X}$ comprises the 48 available discrete values for the areas of the bracing elements. The first failure event is defined in terms of the normalized demand function

$$d_1(\boldsymbol{x}, \boldsymbol{\theta}) = \max_{\upsilon = x, y} \max_{t \in [0,T]} \frac{\left| u_r^{\upsilon}(t, \boldsymbol{x}, \boldsymbol{\theta}) \right|}{u_r^*} \tag{31}$$

with $u_r^{\upsilon}(t, \boldsymbol{x}, \boldsymbol{\theta})$ the displacement at the roof centroid along the x or y direction and $u_r^* = 0.033$ m. Similarly, the normalized demand function corresponding to the second failure event is given by

$$d_2(\boldsymbol{x}, \boldsymbol{\theta}) = \max_{\ell = 1, \ldots, 128} \max_{t \in [0,T]} \frac{|\sigma_\ell(t, \boldsymbol{x}, \boldsymbol{\theta})|}{\sigma^*} \tag{32}$$

where $\sigma_\ell(t, \boldsymbol{x}, \boldsymbol{\theta})$ is the axial stress of the ℓth brace element and $\sigma^* = 3.31 \times 10^8$ Pa (80% of yield stress).

The proposed approach is implemented considering $n = 100$ samples per stage and $\nu = 0.5$. For illustration purposes, the exploration phase considers a target sample size of 250 feasible designs while the exploitation phase stops after twelve stages.

It is noted that, in this case, all design variables are discrete. For the exploration phase the corresponding proposal distribution considers $\lambda_i^* = 5$, $i = 1, \ldots, 4$, and $\tau = 0.025$, while for the exploitation phase $\lambda_i^* = 2$ and $\tau = 0$. Finally, the adaptive surrogate model strategy is not implemented for this example.

The feasible designs obtained from the exploration phase are presented in Fig. 10 in terms of two-dimensional projections and marginal histograms. This set is composed of 250 feasible designs obtained in four stages. The results show that the range of the design variables associated with the lower floors (x_1 and x_2) is smaller than of those corresponding to upper floors (x_3 and x_4). Thus, the system performance seems to be more sensitive to the stiffness of lower floors than of upper floors, which is reasonable from the structural viewpoint. This shows some of the advantages of the proposed approach, in the sense that valuable insight about the system behavior can be obtained as a byproduct of the sampling process.

Starting from the designs in Fig. 10, an exploitation phase is carried out. After ten stages, the set of samples presented in Fig. 11 is obtained. It is seen that these samples densely populate a small portion of the initial search space, i.e., they are

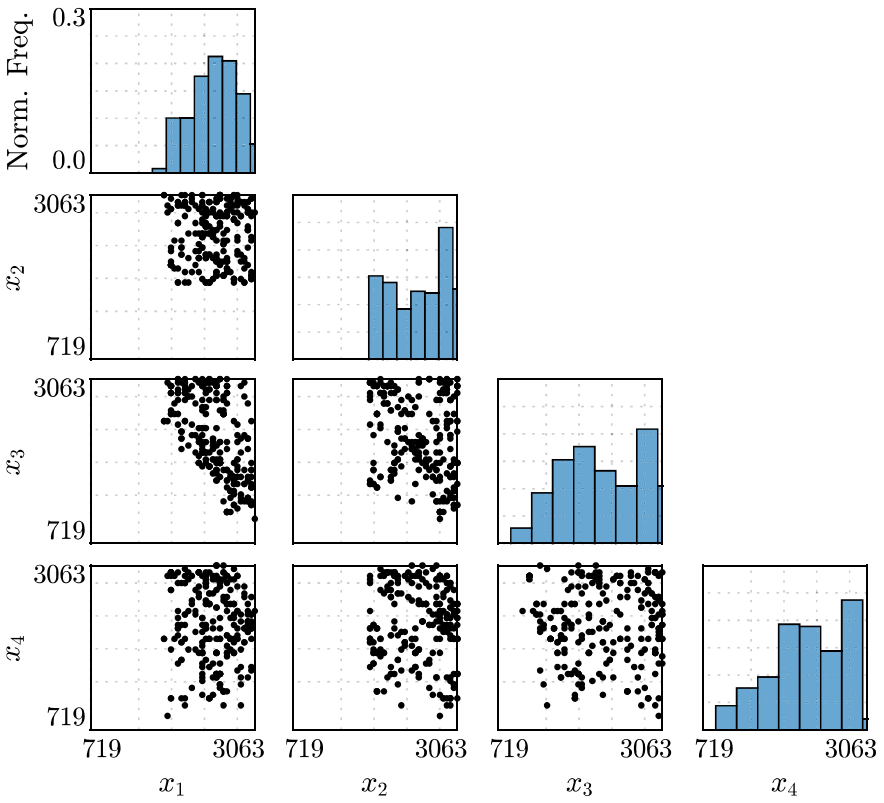


Fig. 10 Set of feasible designs obtained in the exploration phase. Example 3

almost coincident. To obtain further insight into the optimization procedure, Fig. 12 shows the maximum and minimum values of the objective function $f(\boldsymbol{x})$ observed during the different exploitation stages. It is seen that both values are almost coincident from stage $j = 7$ on. The sample-based minimum objective value is equal to 0.6235, which corresponds to $\boldsymbol{x}^T = \langle 2615, 2216, 1418, 1368 \rangle$ mm^2. It is noted that this solution imposes larger cross sections for the lower floors, which is consistent from an engineering viewpoint. The corresponding failure probability values satisfy $P_{F_1}(\boldsymbol{x})/5 \times 10^{-4} = 0.99$ and $P_{F_2}(\boldsymbol{x})/5 \times 10^{-4} = 0.05$. Thus, the first reliability constraint, which is associated with the maximum roof displacement, can be regarded as active at this solution. Finally, in terms of numerical efforts, the total number of designs evaluated in this case is in the order of 1000. That is, the algorithm is capable of exploring the optimal solution set of this RBO problem in an effective manner with a relatively small sample size.

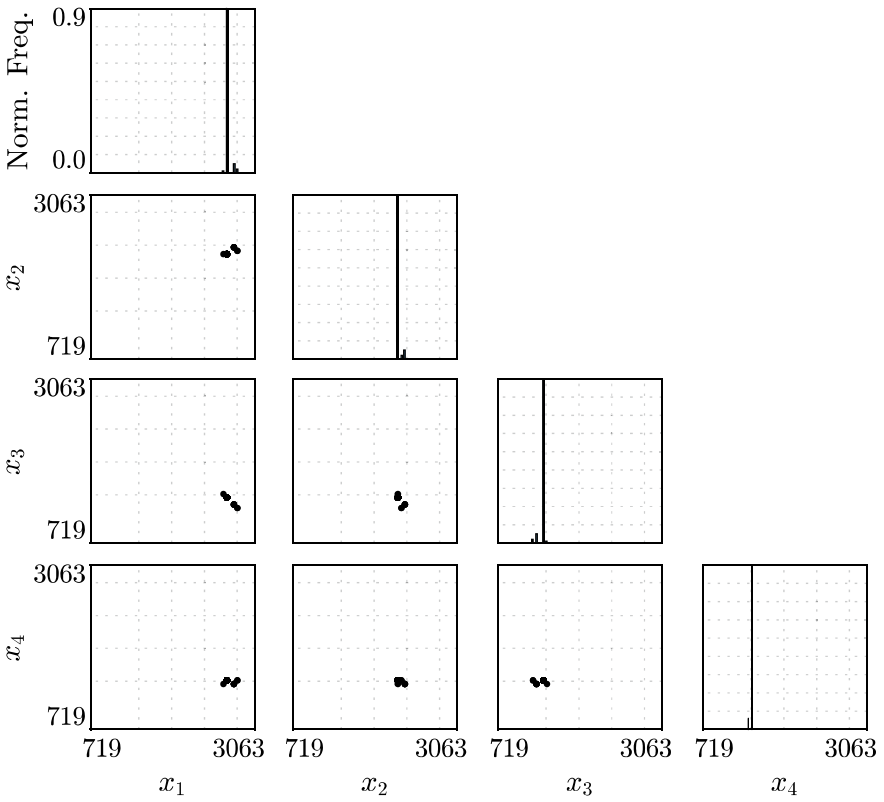


Fig. 11 Set of designs at the end of the exploitation phase. Example 3

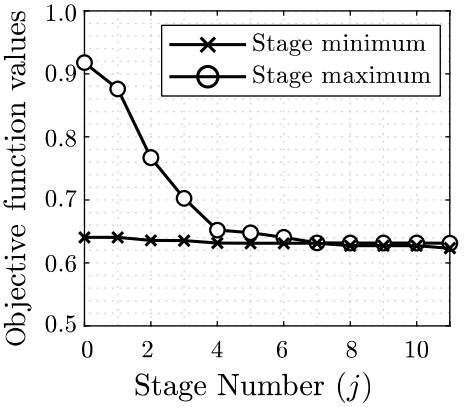


Fig. 12 Maximum and minimum objective function values obtained during the exploitation phase. Example 3

6 Conclusions

A two-phase sampling approach for the reliability-based optimization of structural engineering systems has been presented. The method relies on the reformulation of the constrained optimization problem as obtaining samples uniformly distributed over the optimal solution set. This task is addressed sequentially. An exploration phase is first performed to generate feasible designs, which are then used in an exploitation phase to yield a set of close-to-optimal designs. Due to its theoretical foundations, the method has high chances to reach a vicinity of the optimum solution set. Further, it is relatively simple to implement, it provides flexibility for decision-making processes, and it yields sensitivity information as a byproduct of the sampling process. To illustrate the capabilities of the approach, three examples involving nonlinear structural systems under stochastic ground excitation have been presented, which include continuous and discrete design spaces as well as unconstrained and constrained formulations. Overall, the numerical results indicate that the proposed approach is a potentially useful tool for solving a class of practical RBO problems in structural engineering applications.

References

1. Haftka RT, Gürdal Z (2012) Elements of structural optimization. Springer Science & Business Media. Kluwer, The Netherlands, vol 11
2. Arora J (2004) Introduction to optimum design. Elsevier, United Kingdom
3. Enevoldsen I, Sørensen JD (1994) Reliability-based optimization in structural engineering. Struct Saf 15(3):169–196
4. Gasser M, Schuëller GI (1997) Reliability-based optimization of structural systems. Math Methods Oper Res 46(3):287–307
5. Schuëller GI, Jensen HA (2008) Computational methods in optimization considering uncertainties–an overview. Comput Methods Appl Mech Eng 198(1):2–13

6. Schueller GI, Pradlwarter HJ, Koutsourelakis PS (2004) A critical appraisal of reliability estimation procedures for high dimensions. Probab Eng Mech 19(4):463–474
7. Goller B, Pradlwarter HJ, Schuëller GI (2013) Reliability assessment in structural dynamics. J Sound Vib 332(10):2488–2499
8. Jerez DJ, Jensen HA, Beer M (2022) Reliability-based design optimization of structural systems under stochastic excitation: an overview. Mech Syst Signal Process 166:108397
9. Jensen HA, Ferre MS, Kusanovic DS (2010) Reliability-based synthesis of non-linear stochastic dynamical systems: a global approximation approach. Int J Reliab Saf 4(2–3):139–165
10. Jensen HA, Becerra LG, Valdebenito MA (2013) On the use of a class of interior point algorithms in stochastic structural optimization. Comput Struct 126:69–85
11. Liu WS, Cheung SH (2017) Reliability based design optimization with approximate failure probability function in partitioned design space. Reliab Eng Syst Saf 167:602–611
12. Yuan X, Liu S, Valdebenito MA, Faes MG, Jerez DJ, Jensen HA, Beer M (2021) Decoupled reliability-based optimization using Markov chain Monte Carlo in augmented space. Adv Eng Softw 157:103020
13. Jensen HA, Jerez DJ, Valdebenito M (2020) An adaptive scheme for reliability-based global design optimization: a Markov chain Monte Carlo approach. Mech Syst Signal Process 143:106836
14. Jensen H, Jerez D, Beer M (2021) A general two-phase Markov chain Monte Carlo approach for constrained design optimization: application to stochastic structural optimization. Comput Methods Appl Mech Eng 373:113487
15. Jensen HA, Jerez DJ, Beer M (2022) Structural synthesis considering mixed discrete–continuous design variables: A Bayesian framework. Mech Syst Signal Process 162:108042
16. Jerez DJ, Jensen HA, Beer M, Chen J (2022) Asymptotic Bayesian optimization: a Markov sampling-based framework for design optimization. Probab Eng Mech 67:103178
17. Angelikopoulos P, Papadimitriou C, Koumoutsakos P (2015) X-TMCMC: adaptive kriging for Bayesian inverse modeling. Comput Methods Appl Mech Eng 289:409–428
18. Spence SM, Kareem A (2014) Performance-based design and optimization of uncertain wind-excited dynamic building systems. Eng Struct 78:133–144
19. Bobby S, Spence SM, Bernardini E, Kareem A (2014) Performance-based topology optimization for wind-excited tall buildings: a framework. Eng Struct 74:242–255
20. Petromichelakis I, Psaros A, Kougioumtzoglou I (2021) Stochastic response analysis and reliability-based design optimization of nonlinear electrome-chanical energy harvesters with fractional derivative elements. ASCE-ASME J Risk Uncert Eng Sys Part B Mech Eng 7:1
21. Der Kiureghian A (2000) The geometry of random vibrations and solutions by FORM and SORM. Probab Eng Mech 15(1):81–90
22. Schuëller GI, Pradlwarter HJ (2007) Benchmark study on reliability estimation in higher dimensions of structural systems—an overview. Struct Saf 29(3):167–182
23. Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. Science 220:671–680
24. Černý V (1985) Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. J Optim Theory Appl 45(1):41–51
25. Ching J, Chen YC (2007) Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. J Eng Mech 7:816–832
26. Beck JL, Au SK (2002) Bayesian updating of structural models and reliability using Markov chain Monte Carlo simulation. J Eng Mech 128(4):380–391
27. Wang J, Katafygiotis LS (2014) Reliability-based optimal design of linear structures subjected to stochastic excitations. Struct Saf 47:29–38
28. Zuev KM, Beck JL (2013) Global optimization using the asymptotically independent Markov sampling method. Comput Struct 126:107–119
29. Kanjilal O, Papaioannou I, Straub D (2021) Cross entropy-based importance sampling for first-passage probability estimation of randomly excited linear structures with parameter uncertainty. Struct Saf 91:102090

30. Jensen HA, Vergara C, Papadimitriou C, Millas E (2013) The use of updated robust reliability measures in stochastic dynamical systems. Comput Methods Appl Mech Eng 267:293–317
31. Chatzi EN, Papadimitriou C (eds) (2016) Identification methods for structural health monitoring. Springer, Berlin, vol 567
32. Beck JL (2010) Bayesian system identification based on probability logic. Struct Control Health Monit 17(7):825–847
33. Yuen KV (2010) Bayesian methods for structural dynamics and civil engineering. Wiley, Singapore
34. Betz W, Papaioannou I, Straub D (2016) Transitional Markov chain Monte Carlo: observations and improvements. J Eng Mechan 142(Issue-5)
35. Angelikopoulos P, Papadimitriou C, Koumoutsakos P (2012) Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework. J Chem Phys 137(14):144103
36. Lophaven SN, Nielsen HB, Søndergaard J (2002) DACE: a Matlab kriging toolbox. In: IMM, informatics and mathematical modelling, The Technical University of Denmark, Lyngby, Denmark, vol 2
37. Santner TJ, Williams BJ, Notz WI, Williams BJ (2003) The design and analysis of computer experiments, vol 1. Springer, New York
38. Au SK, Beck JL (2001) Estimation of small failure probabilities in high dimensions by subset simulation. Probab Eng Mech 16(4):263–277
39. Au SK, Wang Y (2014) Engineering risk assessment with subset simulation. Wiley, Singapore
40. Jensen HA, Mayorga F, Papadimitriou C (2015) Reliability sensitivity analysis of stochastic finite element models. Comput Methods Appl Mech Eng 296:327–351
41. Mosqueda G, Whittaker AS, Fenves GL (2004) Characterization and modeling of friction pendulum bearings subjected to multiple components of excitation. J Struct Eng 130(3):433–442
42. Lomiento G, Bonessio N, Benzoni G (2013) Friction model for sliding bearings under seismic excitation. J Earthquake Eng 17(8):1162–1191
43. Boore DM (2003) Simulation of ground motion using the stochastic method. Pure Appl Geophys 160(3):635–676
44. Atkinson GM, Silva W (2000) Stochastic modeling of California ground motions. Bull Seismol Soc Am 90(2):255–274
45. Anderson JG, Hough SE (1984) A model for the shape of the Fourier amplitude spectrum of acceleration at high frequencies. Bull Seismol Soc Am 74(5):1969–1993
46. Mavroeidis GP, Papageorgiou AS (2003) A mathematical representation of near-fault ground motions. Bull Seismol Soc Am 93(3):1099–1131
47. Pradlwarter HJ, Schuëller GI (1993) Equivalent linearization—a suitable tool for analyzing MDOF-systems. Probab Eng Mech 8(2):115–126
48. AISC (American Institute of Steel Construction) (2001) Manual of steel construction load and resistance factor design. AISC, Chicago

Moment Estimation-Based Method of Motion Accuracy Reliability Analysis for Industrial Robots

Dequan Zhang, Shuoshuo Shen, and Xu Han

Abstract Comprehensive and effective assessment of motion accuracy reliability for industrial robot registers a crucial and lasting challenge. In order to ensure the precision performance of industrial robots, this study systematically investigates the reliability modeling and analysis. For kinematic accuracy reliability, a novel computational framework is proposed to comprehensively evaluate the reliability for kinematic positioning and trajectory accuracy of industrial robots, in which the motion error correlation quantification methods are developed. In terms of dynamics accuracy reliability, the rotational sparse grid method and the advanced mixed-degree cubature formula are inferred to evaluate statistical moments of industrial robots' joint torque subject to multidimensional correlations among uncertain parameters. The computational performance of proposed methods is significantly improved compared to the traditional competitive approaches. The engineering practicability and proficiency of the proposed methods are verified by a series of industrial robot examples.

Keywords Industrial robot · Kinematics · Dynamics · Reliability analysis method · Precision performance

1 Introduction

Industrial robot, as electromechanical digital equipment, has become an important indicator of the automation level for manufacturing science, technology, and their high reliability and precision performance are the linchpin to ensure the production efficiency of high-tech industries [1, 2]. However, it is generally argued that the kinematic and dynamic uncertainties deviate the end effector/manipulator from the specified position to negate operational reliability, thus the quality of the products [3, 4]. Therefore, to ensure working performance of industrial robots, it is imperative

D. Zhang · S. Shen · X. Han (✉)
State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Mechanical Engineering, Hebei University of Technology, Tianjin 300401, P. R. China
e-mail: xhan@hebut.edu.cn

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_3

to comprehensively evaluate the reliability of motion accuracy in terms of kinematic and dynamic uncertainties.

Motion accuracy reliability of industrial robots can be specified as the probability of the motion error that falls below a specified error threshold. Recently, some probabilistic methods, such as sampling-based simulation method [5, 6], analytical approximation method [7–9], surrogate model method [10–12], numerical integration-based method [13, 14], etc. have been applied for reliability analysis of industrial robots' motion accuracy. Specifically, Rao and Bhatti [15] employed Monte Carlo Simulation (MCS) to compute the positioning accuracy reliability of industrial robot. Kim et al. [16] proposed analytical approach-based method in comprehending the kinematic uncertainty parameters of industrial robots to evaluate the positioning accuracy reliability subject to normal random variables. Pandey and Zhang [17] put forward a fractional moment-based method pertaining to the extreme value distribution method for reliability analysis of industrial robot. Wu et al. [18] established a new point estimation method to evaluate positioning accuracy reliability. Out of the context of the aforementioned conventional approaches, Lara-Molina et al. [19] explored a novel method for analyzing the accuracy reliability of manipulators using fuzzy theory by considering the uncertainty of joint clearance. Zhao et al. [20] suggested the position error model of manipulators based upon group theory and analyzed the time-dependent motion accuracy reliability for industrial robots. Zhang et al. [21] developed the moment and copula function-based trajectory accuracy reliability analysis method, which can effectively quantify intercorrelations among motion errors. Yang and Yang [22] applied the hierarchical modularization method to assess the kinematic reliability of industrial robots. Cao et al. [23] analyzed the effects of epistemic uncertainty and correlation on the accuracy of manipulator position, and effectively calculated the reliability of positioning accuracy based upon evidence theory. Chen et al. [24] established a robotic dynamic reliability model, and calculated its accuracy reliability via probability and interval hybrid method. Wu et al. [25] developed an improved numerical integration method to approach the industrial robots dynamics reliability. In summary, the motion accuracy reliability analysis constitutes of kinematic and dynamic accuracy reliability. The mathematical model of the kinematic and dynamic accuracy analyses for industrial robots could be highly complex. The traditional sample-based method requires a large number of function calls and simulation time, and the approximate analytical method may lead to unsatisfactory calculation results. Thus, an innovative method to analyze efficiently and reliably the kinematics and dynamics reliabilities of industrial robots are yet to be developed.

Numerical integration-based methods have been widely studied for their high efficiency and applicability [26]. Dimension-reduction method (DRM) [27] is an efficient method to assess multi-dimensional integrals, and the most popular DRM is the bivariate dimension reduction method (BDRM). However, for high dimensional problems, the computational intensity of DRM is usually intractable. Other numerical integral techniques, such as cubature formula [13, 26], sparse grid method [28, 29] and point estimation method [30], have also been employed for acquiring statistical data of system. To analyze proficiently and efficiently the reliability of industrial

robot kinematics and dynamics efficiently, moment-based methods should be further advanced to explore the uncertainty transfer mechanism.

Complex correlation for internal systems and input variables often exhibits in industrial robots. This correlation may influence the system reliability [31]. Therefore, it is important to consider the correlation in estimating failure probability. Copula functions are rigorous and thus widely used in modeling the correlations among mechanical structure parameters and multiple failure modes. In this respect, Lu et al. [32] proposed a system reliability analysis method via combining the moment-matching method with copula function. Wang et al. [33] applied copula function to implement the performance correlation modeling for dynamic systems. Furthermore, the vine copula functions have been engaged for representation of the multidimensional correlations [34, 35]. For example, Li et al. [36] combined the vine copula function and Bayes theorem to develop an estimation approach for structural reliability. Jiang et al. [37] integrated vine copula and first-order reliability method to establish a new reliability analysis method considering inter-dependent random variables. In comprehension of the advantages of the above methods, this study applies copula functions to approach the inter-correlation of random variables or internal robotic system for motion accuracy reliability analysis.

The novelty of this study lies in a new computational framework for evaluating the kinematic accuracy reliability of industrial robots by integrating numerical integration-based methods and probability distribution fitting methods. The eigen-decomposition method and copula functions are developed to quantify the correlation of motion errors. The single coordinate, single point, multiple points, and the trajectory accuracy reliabilities are then comprehensively calculated. A rotational sparse grid method and an advanced mixed-degree cubature formula are constructed to efficiently calculate the reliability of dynamic performance.

Following this introduction, the kinematic computational framework and the engineering example implementation are presented in Sect. 2. Two moment-based methods and one practical dynamic example are exposited in Sect. 3. Conclusions are drawn in Sect. 4.

2 Kinematic Accuracy Reliability Analysis

2.1 Problem Statement

The position of an industrial robot in actual operation can be expressed as

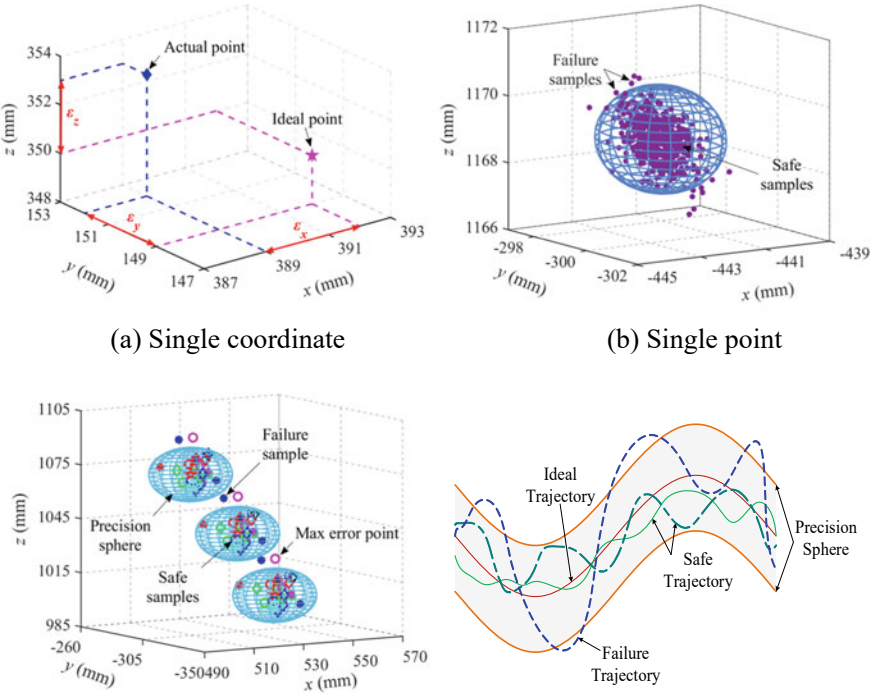
$$\mathbf{p}_a = D(\mathbf{e}) \tag{1}$$

where $\mathbf{p}_a = [p_{ax}, p_{ay}, p_{az}]^{\mathrm{T}}$ denotes the actual position; $D(\cdot)$ represents the kinematic equation; $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]^{\mathrm{T}}$ stands for the random variables; $\mathbf{e}_1 = [a_1, d_1, \theta_1, \alpha_1]$ is the random variables of kinematic parameters.

Due to the uncertainties in input variables, the error of the actual position from the ideal position of industrial robot's exhibit. When the positioning error of the industrial robot in the three coordinate directions is greater than a specified accuracy threshold, the single coordinate (SC) positioning performance would be considered as a failure. The single point (SP) positioning performance reflects the probability that the positional error of point falls below the specified threshold. The multi-point (MP) positioning accuracy reliability indicates the probability that the bounding value of errors for multiple actual positioning points during its spatial motion stabilize within the specific error threshold. Industrial robot trajectory accuracy reliability specifies the probability that the actual motion trajectory stays within the specified trajectory error range [29]. The failure criterions of single coordinate positioning, single point positioning, multi point positioning and trajectory are illustrated in Fig. 1.

Single coordinate positioning accuracy: Denoting the ideal position as $\mathbf{p}_d = [p_{dx}, p_{dy}, p_{dz}]^{\mathrm{T}}$, the error $\mathbf{p}_\mu$ can be specified as [18]

$$\mathbf{p}_\mu = \mathbf{p}_a - \mathbf{p}_d = [\varepsilon_x(\mathbf{e}), \varepsilon_y(\mathbf{e}), \varepsilon_z(\mathbf{e})]^{\mathrm{T}} \tag{2}$$



(a) Single coordinate (b) Single point

(c) Multi point (d) Trajectory

Fig. 1 Failure criteria for kinematic accuracy

where $\varepsilon_x(\mathbf{e})$, $\varepsilon_y(\mathbf{e})$ and $\varepsilon_z(\mathbf{e})$ indicate functions with respect to $\mathbf{e}$, which represent the deviation between the ideal and the actual positions in the three coordinate directions, respectively. The specific terms in Eq. (2) can be derived as

$$\varepsilon_x(\mathbf{e}) = |p_{ax} - p_{dx}|,\ \varepsilon_y(\mathbf{e}) = \left|p_{ay} - p_{dy}\right|,\ \varepsilon_z(\mathbf{e}) = |p_{az} - p_{dz}| \tag{3}$$

The performance function in three coordinate directions can be described as

$$G_x(\mathbf{e}) = \varepsilon_x(\mathbf{e}) - r_x,\ G_y(\mathbf{e}) = \varepsilon_y(\mathbf{e}) - r_y,\ G_z(\mathbf{e}) = \varepsilon_z(\mathbf{e}) - r_z \tag{4}$$

where r_x, r_y and r_z denote the accuracy thresholds in three coordinate directions. When the probability density functions (PDFs) of $\varepsilon_x(\mathbf{e})$, $\varepsilon_y(\mathbf{e})$ and $\varepsilon_z(\mathbf{e})$ are available, the positioning accuracy reliability $R_{(\cdot)}$ and failure probability $P_{f(\cdot)}$ can be expressed as

$$\begin{cases} R_x = \Pr[G_x(\mathbf{e}) < 0] = \int\limits_{G_x(\mathbf{e})<0} f_{ex}(\mathbf{e})\mathrm{d}\mathbf{e},\ P_{fx} = 1 - R_x \\ R_y = \Pr\left[G_y(\mathbf{e}) < 0\right] = \int\limits_{G_y(\mathbf{e})<0} f_{ey}(\mathbf{e})\mathrm{d}\mathbf{e},\ P_{fy} = 1 - R_y \\ R_z = \Pr\left[G_z(\mathbf{e}) < 0\right] = \int\limits_{G_z(\mathbf{e})<0} f_{ez}(\mathbf{e})\mathrm{d}\mathbf{e},\ P_{fz} = 1 - R_z \end{cases} \tag{5}$$

where $f_{ex}(\mathbf{e})$, $f_{ey}(\mathbf{e})$ and $f_{ez}(\mathbf{e})$ represent the probability density function of positioning error in x, y and z coordinate directions, respectively.

Single point positioning accuracy: The distance between the ideal and the actual positions $\varepsilon(\mathbf{e})$ can be formulated as [2]

$$\varepsilon(\mathbf{e}) = \sqrt{\varepsilon_x^2(\mathbf{e}) + \varepsilon_y^2(\mathbf{e}) + \varepsilon_z^2(\mathbf{e})} \tag{6}$$

The performance function can be written as

$$G_p(\mathbf{e}) = \varepsilon(\mathbf{e}) - r_s = \sqrt{\varepsilon_x^2(\mathbf{e}) + \varepsilon_y^2(\mathbf{e}) + \varepsilon_z^2(\mathbf{e})} - r_s \tag{7}$$

where r_s indicates the accuracy threshold. Reliability R_p and failure probability P_{fp} are then calculated by

$$R_p = \Pr\left[G_p(\mathbf{e}) < 0\right] = \int\limits_{G_p(\mathbf{e})<0} f_e(\mathbf{e})\mathrm{d}\mathbf{e},\ P_{fp} = 1 - R_p \tag{8}$$

where $f_e(\mathbf{e})$ stands for the joint PDF.

Multi point positioning accuracy: The error extreme value $\varepsilon_{\max}(\mathbf{e})$ at multiple positioning points of an industrial robot indicates its multi-point positioning accuracy, which takes the form of [29]

$$\varepsilon_{\max}(\mathbf{e}) = \max\left[\varepsilon_1(\mathbf{e}), \varepsilon_2(\mathbf{e}), \ldots, \varepsilon_j(\mathbf{e}), \ldots, \varepsilon_n(\mathbf{e})\right], j = 1, 2, \ldots, n \tag{9}$$

where $\varepsilon_j(\mathbf{e})$ represents the distance between the ideal and the actual positional points; n symbolizes the number of positional points. This performance function can be derived by

$$G_{mp}(\mathbf{e}) = \varepsilon_{\max}(\mathbf{e}) - r_{ms} \tag{10}$$

where r_{ms} indicates the accuracy threshold of multi-point. Reliability R_{mp} and failure probability P_{fmp} can be obtained as

$$R_{mp} = \Pr\left[G_{mp}(\mathbf{e}) < 0\right] = \int\limits_{G_{mp}(\mathbf{e})<0} f_e(\mathbf{e})\mathrm{d}\mathbf{e},\ P_{fmp} = 1 - R_{mp} \tag{11}$$

Trajectory accuracy reliability: For the trajectory accuracy reliability of industrial robots, it is prerequisite to assess the coupling of discrete point positioning failure [21]. If the trajectory of industrial robot is discretized into n discrete positioning points, the trajectory accuracy reliability can be formulated as

$$\begin{aligned} R_T &= \mathrm{Prob}[G_1(\mathbf{e}) < 0 \cap G_2(\mathbf{e}) < 0 \cap \ldots \cap G_j(\mathbf{e}) < 0 \ldots \cap G_n(\mathbf{e}) < 0] \\ j &= (1,\ 2, \ldots, n) \end{aligned} \tag{12}$$

The failure probability is derived by

$$P_{fT} = 1 - R_T \tag{13}$$

2.2 Kinematic Reliability Analysis Method

The positioning errors of three coordinates are incurred by uncertainties in variables and there exhibit approximately linear correlations among three coordinates errors. Assuming positioning error as Gaussian distribution, the square of the single point positioning error can be approximated by the eigen-decomposition method [21]:

$$\chi^2(\mathbf{e}) = \varepsilon^2(\mathbf{e}) = \sum_{i=1}^{3} \lambda_i(\alpha_i + u_i)^2 \tag{14}$$

where λ_1, λ_2 and λ_3 are eigenvalues of covariance matrix among three coordinate directional errors; $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$ represents vector of the undetermined coefficient; $\mathbf{u} = (u_1, u_2, u_3)^{\mathrm{T}}$ denotes independent standard normal variable.

For the multi-points positioning error, the error of each positioning point is derived first, and the multi-points positioning error is then converted using the extreme value theory. The conversion of the extreme value error can be expressed as

$$\chi^2_{\max}(\mathbf{e}) = \max[\chi^2_1(\mathbf{e}), \ldots, \chi^2_j(\mathbf{e}), \ldots, \chi^2_n(\mathbf{e})] \tag{15}$$

2.2.1 Positioning Accuracy Reliability Analysis Method

The i-order origin moments M_i of $Y(\mathbf{e})$ are defined as [13]

$$M_i = \int_{\Omega} Y^i(\mathbf{e}) f_e(\mathbf{e})\, \mathrm{d}\mathbf{e}, i = 1, 2, \ldots$$

$$\text{SC: } Y(\mathbf{e}) = \varepsilon_x(\mathbf{e}), \varepsilon_y(\mathbf{e}) \text{ or } \varepsilon_z(\mathbf{e});\ \text{SP} : Y(\mathbf{e}) = \chi(\mathbf{e});\ \text{MP} : Y(\mathbf{e}) = \chi_{\max}(\mathbf{e}) \tag{16}$$

where Ω is the uncertainty distribution domain.

Through sparse grid (SPGR) quadrature rule, the multi-dimensional node $\mathbf{S}_n^k$ can be obtained as follows [38],

$$\begin{aligned} \mathbf{S}_n^k &= \bigcup_{k+1 \le |m| \le k+n} X_1^{m_1} \otimes X_1^{m_2} \otimes \ldots \otimes X_1^{m_n} \\ &= \bigcup_{k+1 \le |m| \le k+n} T^{-1}(U_1^{m_1}) \otimes T^{-1}(U_1^{m_2}) \otimes \ldots \otimes T^{-1}(U_1^{m_n}) \end{aligned} \tag{17}$$

where m_i ($i = 1, 2, \ldots, n$) is univariate indicator; $U_1^{m_1}$ represents Gauss-Hermite integration points; $X_1^{m_1}$ denotes the integration points to the input random variables and $X_1^{m_1} = T^{-1}(U_1^{m_1})$; $T^{-1}(\cdot)$ stands for the inverse Rosenblatt transformation; $\otimes$ is the tensor operation that is defined as

$$\begin{cases} \text{if} : X_1 = [a, b]^{\mathrm{T}}, X_2 = [1, 2]^{\mathrm{T}} \\ \text{then} : X_1 \otimes X_2 = \{(a, 1), (a, 2), (b, 1), (b, 2)\} \end{cases} \tag{18}$$

The corresponding weights for each multidimensional node can be inferred as [28]

$$\omega_q = (-1)^{k+n-|m|} \binom{n-1}{k+n-|m|} \omega_{1,j_1}^{m_1} \omega_{1,j_2}^{m_2} \cdots \omega_{1,j_n}^{m_n} \tag{19}$$

where ω_q represents the weight coefficient to the qth multidimensional point $\boldsymbol{\varsigma}_q = [\varsigma_{1,j_1}^{m_1}, \varsigma_{1,j_2}^{m_2}, \ldots, \varsigma_{1,j_n}^{m_n}]^{\mathrm{T}} \in \mathbf{S}_n^k$; $\omega_{1,j}^{m_1}$ stands for the weight of the one-dimensional

point $\varsigma_{1,j}^{m_1}$. According to the SPGR quadrature rule, Eq. (16) can be approached by

$$M_i = \int_{\Omega} Y^i(\mathbf{e}) f_e(\mathbf{e})\, \mathrm{d}\mathbf{e} \approx \sum_{q=1}^{N} \omega_q Y^i(\varsigma_q),\ i = 1, 2, \ldots \tag{20}$$

Saddlepoint approximation (SPA) technique is applied herein to evaluate the reliability of positioning accuracy. The cumulative generating function can be expressed as [39]

$$K_Y(t) = \ln(M_Y(t)) = \ln\left[\int_{-\infty}^{\infty} e^{ty} f_Y(y)\mathrm{d}y\right] \tag{21}$$

For convenience, Eq. (21) takes up expansion form:

$$K_Y(t) = k_1 + \frac{k_2 t^2}{2!} + \frac{k_3 t^3}{3!} + \frac{k_4 t^4}{4!} \tag{22}$$

It first two order derivative expressions can be written as

$$K_Y'(t) = k_1 + \sum_{s=2}^{4} k_s \frac{t^{s-1}}{(s-1)!},\ K_Y''(t) = k_2 + \sum_{s=3}^{4} k_s \frac{t^{s-2}}{(s-2)!} \tag{23}$$

where k_1, k_2, k_3 and k_4 can be obtained as

$$\begin{cases} k_1 = M_1, k_2 = M_2 - M_1^2, k_3 = M_3 - 3M_2M_1 + 2M_1^3 \\ k_4 = M_4 - 4M_3M_1 - 3M_2^2 + 12M_2M_1^2 - 6M_1^4 \end{cases} \tag{24}$$

Combining with Eqs. (21)–(24), the saddlepoint t_s can be acquired by solving $K_Y'(t) = r_s$.

Positioning accuracy reliability p_R can then be approximated by SPA technique [40]:

$$p_R = F(r^2) = \Phi(w) + \phi(w)\left(\frac{1}{w} - \frac{1}{v}\right) \tag{25}$$

where $\phi(\cdot)$ is the standard normal probability density functions; w and v can be computed as

$$w = \mathrm{sgn}(t_s)\{2[t_s y - K(t_s)]\}^{1/2} \tag{26}$$

$$v = t_s[K''(t_s)]^{1/2} \tag{27}$$

2.2.2 Trajectory Accuracy Reliability Analysis Method

Multi-points failure mode correlation can be approached based upon the optimal copula functions. The specific derivations, which are only briefed herein, are seen in Refs. [31, 32]. The correlation coefficient τ_{ij} between two failure modes can be expressed by [21]

$$\tau_{ij} = 4\int_{-1}^{1}\int_{-1}^{1}\hat{C}\left(u_i, u_j|\vartheta\right)\mathrm{d}\hat{C}\left(u_i, u_j|\vartheta\right) - 1 \tag{28}$$

where $\hat{C}$ denotes the optimal copula functions. For two-point failures, the joint failure probability can be formulated as

$$\tilde{P}_{fij} = \hat{C}\left(P_{fi}^{\text{SPGR-SPA}}, P_{fj}^{\text{SPGR-SPA}}|\vartheta\right) \tag{29}$$

where $P_{fi}^{\text{SPGR-SPA}}$ and $P_{fj}^{\text{SPGR-SPA}}$ stand for the failure probability at different points, respectively.

For simplicity of presentation, two proposed trajectory accuracy reliability analysis methods are denoted as type A and type B, respectively.

Type A: Ditlevsen's bound method is applied to assess the kinematic trajectory accuracy reliability among multiple points [41]:

$$P_{fT} \cong \sum_{i=1}^{M} P_{fi}^{\text{SPGR-SPA}} - \sum_{i=2}^{M}\max_{j<i}\left(\tilde{P}_{fij}\right) \tag{30}$$

Type B: The trajectory accuracy reliability is derived according to the multivariate normal distribution theory [42]:

$$R_T = \Pr[\bigcap_{j=1}^{n}(\chi_j^2(\mathbf{e}) < r_s^2)] = \Phi_n(\mathbf{0}; \boldsymbol{\beta}_Z, \boldsymbol{\Sigma}_Z) = \int_{-\infty}^{0}\cdots\int_{-\infty}^{0} f_Z(\mathbf{z})\mathrm{d}\mathbf{z} \tag{31}$$

where $f_Z(\mathbf{z})$ is the joint probability density function of $\mathbf{Z}$:

$$\begin{cases} \boldsymbol{\beta}_Z = [\Phi^{-1}(P_{f1}^{\text{SPGR-SPA}}), \Phi^{-1}(P_{f2}^{\text{SPGR-SPA}}), \ldots, \Phi^{-1}(P_{fn}^{\text{SPGR-SPA}})] \\ \boldsymbol{\Sigma}_Z = \begin{bmatrix} 1 & \tau_{12} & \cdots & \tau_{1n} \\ \tau_{21} & 1 & \cdots & \tau_{2n} \\ \vdots & \vdots & \tau_{ij} & \vdots \\ \tau_{n1} & \tau_{n2} & \cdots & 1 \end{bmatrix}_{n\times n} \end{cases} \tag{32}$$

With the derived $\boldsymbol{\beta}_Z$ and $\boldsymbol{\Sigma}_Z$, then $f_Z(\mathbf{z})$ is expressed as

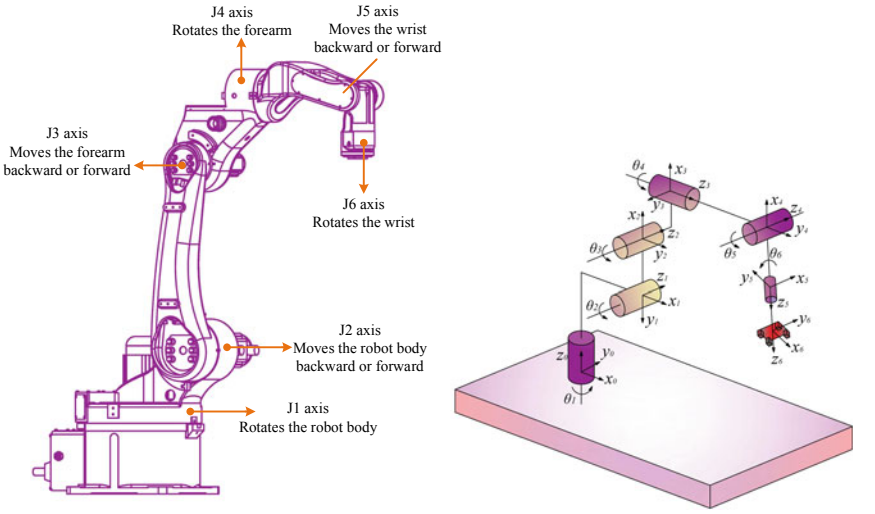
$$f_Z(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_Z|}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\beta}_Z)^{\mathrm{T}} \boldsymbol{\Sigma}_Z^{-1} (\mathbf{z}-\boldsymbol{\beta}_Z)\right) \tag{33}$$

R_T can be calculated by substituting Eq. (33) in Eq. (31), and the failure probability is obtained as

$$P_{fT} = 1 - R_T = 1 - \Phi_n(\mathbf{0}; \boldsymbol{\beta}_Z, \boldsymbol{\Sigma}_Z) \tag{34}$$

2.3 Demonstrative Kinematic Examples

An industrial robot with 6-degree of freedom (6-DoF), as sketched in Fig. 2 [21], is exemplified to demonstrate the superiority of the currently proposed methods for the single coordinate positioning accuracy, single point positioning accuracy, multi-point positioning accuracy and trajectory accuracy reliability analyses. The maximum entropy with fractional moments (ME-FM) [17] and MCS are served as calibers to validate the proposed methods.



(a) Sketch for 6-DoF welding robot (b) Mechanism diagram of 6-DoF welding robot

Fig. 2 A 6-DoF welding robot

Table 1 Kinematic parameters for industrial robot [21]

No.	θ_i	d_i	a_i	α_i
1	θ_1	L_1	L_4	α_1
2	$-30°$	0	L_5	0
3	30°	0	L_6	90°
4	0°	L_2	0	$-90°$
5	θ_5	0	0	α_2
6	0°	L_3	0	0

Table 2 Stats for random variables [21]

Variables	Mean	Standard deviation	Distribution
L_1	463 (mm)	0.463 (mm)	Normal
L_2	581.5 (mm)	0.5815 (mm)	Normal
L_3	203.2 (mm)	0.2032 (mm)	Normal
L_4	220 (mm)	0.220 (mm)	Normal
L_5	600 (mm)	0.600 (mm)	Normal
L_6	220 (mm)	0.220 (mm)	Normal
α_1	90 (°)	0.01 (°)	Normal
α_2	90 (°)	0.01 (°)	Normal

2.3.1 Example 2.1: Single Coordinate

The kinematic parameters and statistical data are summarized in Tables 1 and 2, respectively.

The first-four order origin moments of positioning errors in three coordinate directions are calculated by the proposed method and MCS are summarized in Table 3. Compared with MCS (10^6 runs), the currently proposed method yields more accurate results with only 177 ($k = 2, 2 \times 8^2 + 6 \times 8 + 1 = 177$ runs) and lower maximum error around 0.03%.

PDF of three-coordinate positioning errors and failure probability curves of the industrial robots are shown in Figs. 3 and 4, respectively. The results by the proposed method are more coincident with those by MCS than ME-FM method (500 runs) counterpart. It is verified that the proposed methods can efficiently analyze the reliability of single coordinate positioning accuracy.

2.3.2 Example 2.2: Single Point

Table 4 compares the first four order origin moments of the single point positioning error from the proposed method and MCS. It can be found that the superiority of the proposed method is again verified through this comparison.

Table 3 Statistical moments of positioning errors in x, y and z coordinate directions

Coordinates	Moments	MCS	Proposed method (error %)
X	M_1	0.4829	0.4829 (0.00)
	M_2	0.2372	0.2372 (0.00)
	M_3	0.1185	0.1184 (0.02)
	M_4	0.0601	0.0601 (0.03)
Y	M_1	0.4792	0.4792 (0.00)
	M_2	0.2300	0.2300 (0.00)
	M_3	0.1105	0.1105 (0.00)
	M_4	0.0532	0.0532 (0.00)
Z	M_1	0.2922	0.2922 (0.00)
	M_2	0.0919	0.0919 (0.00)
	M_3	0.0306	0.0306 (0.00)
	M_4	0.0107	0.0107 (0.00)
N_{call}		1×10^6	177

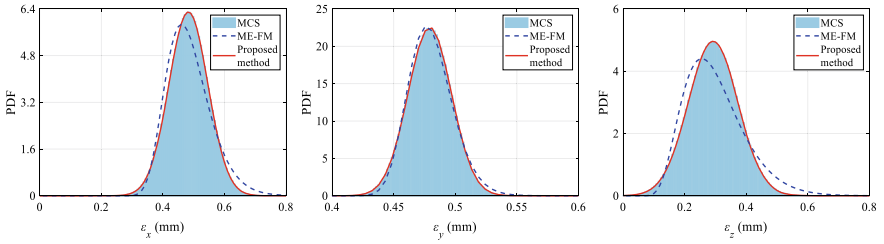


Fig. 3 PDFs of three-coordinate positioning errors

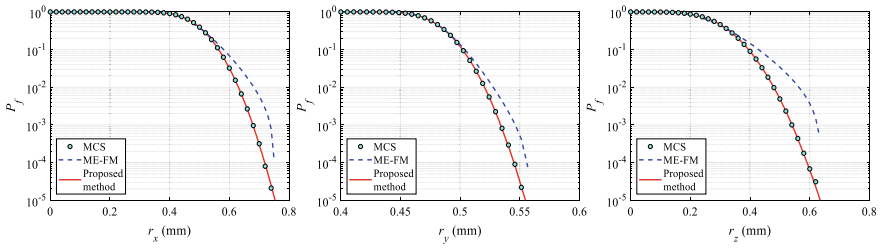
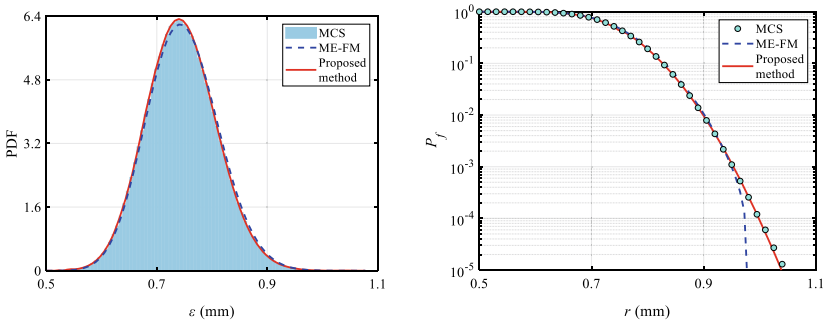


Fig. 4 Failure probabilities of three-coordinate positioning accuracy

Table 4 Statistical moments of single point positioning error

Moments	M_1	M_2	M_3	M_4	Function calls
MCS	0.7450	0.5591	0.4225	0.3216	1×10^6
Proposed method (error %)	0.7450 (0.00%)	0.5591 (0.00%)	0.4225 (0.00%)	0.3216 (0.00%)	177



(a) PDF of single point positioning error (b) P_f of single point positioning accuracy

Fig. 5 Positioning accuracy reliability analysis results for single point

The results of PDF and failure probability are shown in Fig. 5. Compared with ME-FM method, the curves by the proposed methods accord well with those by MCS. Especially in the tail of the failure probability, the curve by ME-FM method deviates significantly from those by MCS, which also verifies the superiority of the proposed method.

2.3.3 Example 2.3: Multi-points

The proposed method is applied to analyze the reliability of multi-points positioning accuracy in this example. Driving θ_1 and θ_5 forms a trajectory, as sketched in Fig. 6. Table 5 presents the driving schemes for θ_1 and θ_5. Interval points 3, 5, 7, 9 and 11 on the trajectory are chosen to verify the effectiveness of the proposed method.

As shown in Table 6, the proposed method can efficiently calculate the first fourth order origin moments with a maximum error as low as 0.26%. It can be stated that

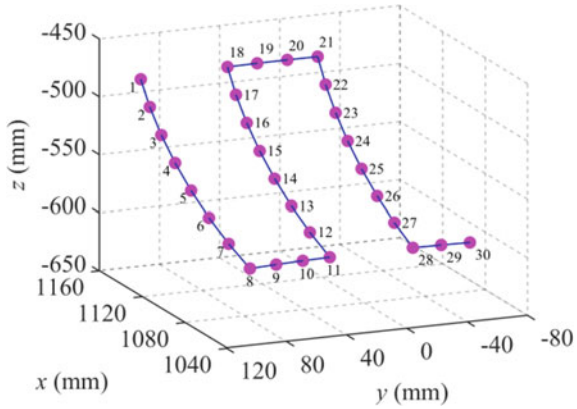


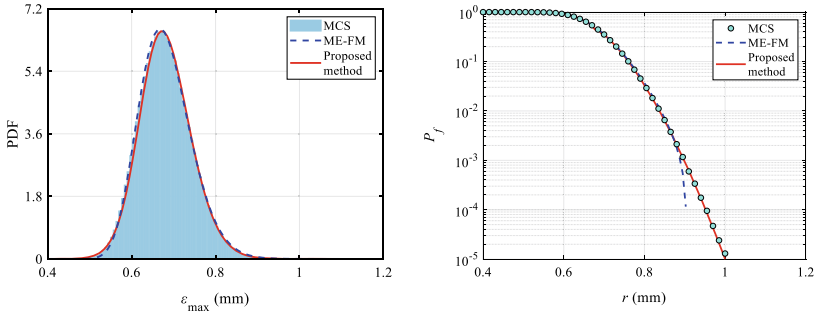
Fig. 6 Schematic for kinematic trajectory [21]

Table 5 Driving schemes of θ_1 and θ_5

Angles	1–8	9–11	12–18	19–21	22–28	29–30
θ_1 (°)	5	[4: 1: 2]	2	[1: − 1: − 1]	− 1	[−2: − 1: − 3]
θ_5 (°)	[72: − 6: 30]	30	[36: 6: 72]	72	[66: − 6: 30]	30

Table 6 Statistical moments for the extremum of multi-point positioning error

Moments	M_1	M_2	M_3	M_4	Function calls
MCS	0.6796	0.4656	0.3217	0.2214	1×10^6
Proposed method (Error %)	0.6801 (0.08%)	0.4664 (0.16%)	0.3224 (0.22%)	0.2247 (0.26%)	177



(a) PDF of multi-points positioning extreme error (b) P_f of multi-points positioning accuracy

Fig. 7 Reliability analysis results of multi-points positioning accuracy

the proposed methods yield remarkably satisfactory result in the reliability analysis of multi-point positioning accuracy.

Figure 7 compares PDF of the multi-point positioning extremum error and failure probability curves by the proposed methods, MCS and ME-FM method. It can be found that significantly approximating results to those by MCS are obtained by the proposed method. A remarkably favorable note is that the currently proposed method is highly more efficient than MCS.

2.3.4 Example 2.4: Trajectory

To further validate the proposed method, analysis for the kinematic trajectory containing 30 discrete points as shown in Fig. 6 is performed to compute the trajectory accuracy reliability. For the 30 discrete points, the nonlinear correlation coefficients can be obtained from Fig. 8.

The failure probability of kinematic trajectory from different methods are summarized in Fig. 9. The proposed methods (A and B) yield better fit to those by MCS

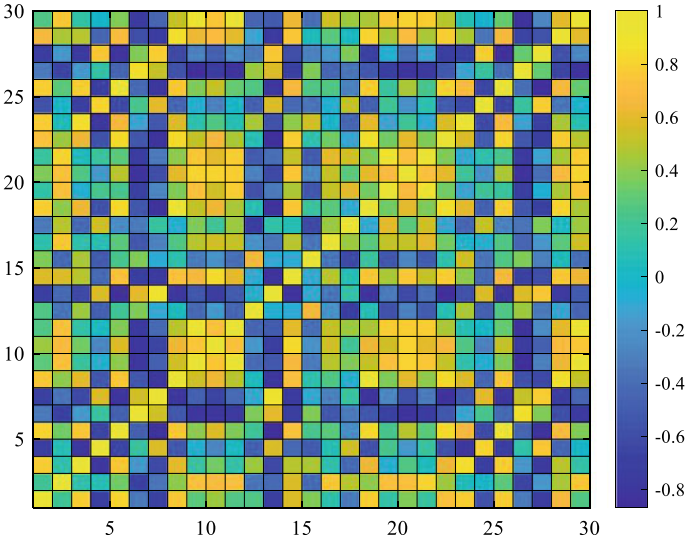


Fig. 8 Correlation coefficients between 30 pairs of discrete points for Example 2.4

than ME-FM method and M-SPA method. Once again, the proposed methods are proved to be highly efficient and reliable in analyzing the reliability of the kinematic trajectory.

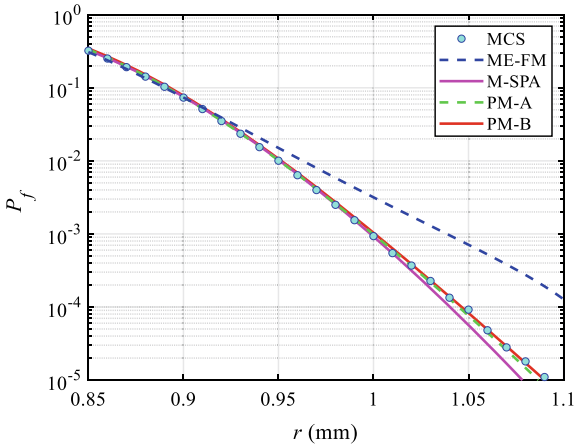


Fig. 9 Failure probability of kinematic trajectory

3 Dynamic Accuracy Reliability Analysis

3.1 Problem Statement

The governing equations of the industrial robot dynamics model can be described as [25]

$$\mathbf{M}(\boldsymbol{\Theta}(t))\ddot{\boldsymbol{\Theta}}(t) + \mathbf{C}(\boldsymbol{\Theta}(t), \dot{\boldsymbol{\Theta}}(t))\dot{\boldsymbol{\Theta}}(t) + \boldsymbol{\tau}_f(t) = \boldsymbol{\tau}(t) \tag{35}$$

where $\mathbf{M}(\cdot)$ represents the mass matrix; $\mathbf{C}(\cdot)$ stands for inertia forces; $\boldsymbol{\tau}_f(t)$ and $\boldsymbol{\tau}(t)$ denote the frictional forces and the joint torque vector, respectively; $\boldsymbol{\Theta}(t)$, $\dot{\boldsymbol{\Theta}}(t)$ and $\ddot{\boldsymbol{\Theta}}(t)$ indicate the angular displacement, velocity and acceleration, respectively.

The dynamics performance function can be derived by [25]

$$\boldsymbol{\tau}(t) = \mathbf{Y}(\boldsymbol{\Theta}(t), \dot{\boldsymbol{\Theta}}(t), \ddot{\boldsymbol{\Theta}}(t))\boldsymbol{\beta} \tag{36}$$

where $\mathbf{Y}(\cdot)$ denotes the regressor matrix; $\boldsymbol{\beta}$ stands for the system parameters, the specific parameters are seen in [25].

The angular position, velocity and acceleration of joint i can be obtained by [43]

$$\begin{cases} \Theta_i(t) = \sum_{l=1}^{N_i} \dfrac{a_l^i}{\omega_f l} \sin(\omega_f l t) - \dfrac{b_l^i}{\omega_f l} \cos(\omega_f l t) + \Theta_{i0} \\ \dot{\Theta}_i(t) = \sum_{l=1}^{N_i} a_l^i \cos(\omega_f l t) + b_l^i \sin(\omega_f l t) \\ \ddot{\Theta}_i(t) = \sum_{l=1}^{N_i} -a_l^i \omega_f l \sin(\omega_f l t) + b_l^i \omega_f l \cos(\omega_f l t) \end{cases} \tag{37}$$

where N_i denotes the number of harmonics; a_l^i and b_l^i indicate Fourier coefficients; ω_f represents the fundamental frequency; Θ_{i0} is the angular displacement offset.

N_c sampling is performed in one trajectory period, Eq. (36) can be reformulated as [43]

$$\begin{cases} \boldsymbol{\Xi} = \boldsymbol{W}\boldsymbol{\beta} \\ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{Y}(\boldsymbol{\Theta}(t_1), \dot{\boldsymbol{\Theta}}(t_1), \ddot{\boldsymbol{\Theta}}(t_1)) \\ \boldsymbol{Y}(\boldsymbol{\Theta}(t_2), \dot{\boldsymbol{\Theta}}(t_2), \ddot{\boldsymbol{\Theta}}(t_2)) \\ \vdots \\ \boldsymbol{Y}(\boldsymbol{\Theta}(t_{N_c}), \dot{\boldsymbol{\Theta}}(t_{N_c}), \ddot{\boldsymbol{\Theta}}(t_{N_c})) \end{bmatrix} \end{cases} \tag{38}$$

The parameters a_l^i, b_l^i and Θ_{i0} in Eq. (37) can be obtained by solving the following optimization model,

$$\begin{cases} \text{Find: } f = \dfrac{1}{\prod\limits_{g=1}^{N}\left(\sum\limits_{k=1}^{6N_c} W_{kg}^2\right)} \\ \text{s.t.:} \\ |\Theta_i(t)| \le \Theta_{i-\max},\ \left|\dot{\Theta}_i(t)\right| \le \dot{\Theta}_{i-\max},\ \left|\ddot{\Theta}_i(t)\right| \le \ddot{\Theta}_{i-\max} \\ \Theta_i(t_0) = \Theta_i(t_{\text{end}}) = 0,\ \dot{\Theta}_i(t_0) = \dot{\Theta}_i(t_{\text{end}}) = 0,\ \ddot{\Theta}_i(t_0) = \ddot{\Theta}_i(t_{\text{end}}) = 0 \end{cases} \tag{39}$$

where W_{kg} represents the element in the k-th row and g-th column of the matrix W; $\Theta_{i\text{ - max}}$, $\dot{\Theta}_{i\text{ - max}}$ and $\ddot{\Theta}_{i\text{ - max}}$ denote the maximum values of the angular displacement, angular velocity and angular acceleration of the i-th joint, respectively; t_0 and t_{end} stand for the initial and final sampling moments, respectively.

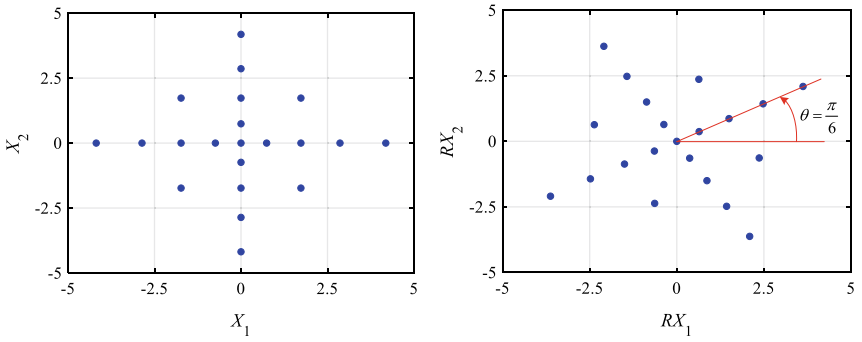
Due to measurement noise and parameter identification errors, the identified dynamic parameters of industrial robots usually exhibit uncertainties to negate the joint torque precision of industrial robots. To analyze the reliability of joint torque accuracy subject to the influence of uncertain parameters, some relatable system parameters are assumed to be uncertain parameters.

3.2 Dynamic Reliability Analysis Method

In this section, the rotational sparse grid method is proposed to efficiently evaluate the statistical moment for dynamic performance response. For dynamic uncertainty parameter correlation problem, an advanced mixed-degree cubature formula is developed to assess the reliability of dynamic performance by integrating vine copula function and Hermite polynomial model.

3.2.1 Rotational Sparse Grid Method

For uncertain parameters with complex distribution types in robot dynamics systems, a rotational sparse grid integration points strategy is developed. This strategy can enlarge the projection rate in the uncertainty space to better capture the distribution information of random variables. In the two-dimensional standard normal space, the original and rotational ($\pi/6$) sparse grid integration points are shown in Fig. 10. Form Fig. 10(a), some integration points coincide in X_1 or X_2 axes projection, and the projection rate is only 3/7. However, the rotational integration points ensue a projection rate of 1 on two coordinate axes in Fig. 10(b), which can capture sufficient probability distribution information.



(a) Original sparse grid points (b) Rotational sparse grid points

Fig. 10 Sparse grid points in 2-dimensional space

Following the above description, the rotating SPGR integration points can better convey information about the probability distribution of random variables. For a 2-dimensional (k, l) plane in an n-dimensional random space, the rotational matrix $\mathbf{R}_{k,l}(\theta)$ can be expressed as [44]

$$\mathbf{R}_{k,l}(\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cos(\theta) & \cdots & -\sin(\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ k\text{-th row} \\ \\ l\text{-th row} \\ \\ n \times n \end{matrix} \tag{40}$$

$$\begin{matrix} & & k\text{-th} & & l\text{-th} & & \\ & & \text{column} & & \text{column} & & \end{matrix}$$

Accordingly, the rotational matrix of all elements can be derived as

$$\mathbf{MR}(\boldsymbol{\theta}) = \prod_{k=1}^{n} \prod_{l=k+1}^{n} \mathbf{R}_{k,l}(\theta_{k,l}) \tag{41}$$

where $\boldsymbol{\theta} = [\theta_{1,2}, \theta_{1,3}, \ldots, \theta_{n-1,n}]^{\mathrm{T}}$.

The n-th multi-dimensional node $\boldsymbol{\varsigma}_q$ is rotated as $\tilde{\boldsymbol{\varsigma}}_q$, such that

$$\tilde{\boldsymbol{\varsigma}}_q = \mathbf{MR}(\boldsymbol{\theta})\boldsymbol{\varsigma}_q = [\tilde{\varsigma}_{q,1}, \tilde{\varsigma}_{q,2}, \ldots, \tilde{\varsigma}_{q,k}, \tilde{\varsigma}_{q,l}, \ldots, \tilde{\varsigma}_{q,n}] \tag{42}$$

Therefore, the rotated n-dimensional node matrix $\tilde{\boldsymbol{\varsigma}}$ can be obtained as

$$\tilde{\boldsymbol{\varsigma}} = [\tilde{\boldsymbol{\varsigma}}_1, \tilde{\boldsymbol{\varsigma}}_2, \ldots, \tilde{\boldsymbol{\varsigma}}_q, \ldots, \tilde{\boldsymbol{\varsigma}}_n]^{\mathrm{T}} \tag{43}$$

The corresponding weights are consistent with the original SPGR. To obtain the optimal integrational points, an optimal rotational angle vector $\boldsymbol{\theta}$ can be derived as follows.

According to the SPGR quadrature rule, statistical moments information of performance function and random variables, an objective function can be established. The statistical moments of performance function with high accuracy are obtained by minimizing the objective function.

With the rotational multi-dimensional points and corresponding weights, the statistical moments of performance function can be obtained by [25]

$$\begin{aligned} M_i &= \sum_{q=1}^{N} \omega_q G^i\left(\tilde{\boldsymbol{\varsigma}}_q\right) = \sum_{q=1}^{N} \omega_q G^i\left(\mathbf{M}\mathbf{R}(\boldsymbol{\theta})\boldsymbol{\varsigma}_q\right) \\ &= \sum_{q=1}^{N} \omega_q G^i\left(\prod_{k=1}^{n}\prod_{l=k+1}^{n} \mathbf{R}_{k,l}(\theta_{k,l})\boldsymbol{\varsigma}_q\right), i = 1, 2, \ldots \end{aligned} \tag{44}$$

where $G(\cdot)$ represents performance function. The marginal moments are formulated as

$$M_{i,X_t} = \sum_{q=1}^{N} \omega_q G^i\left(0, 0, \ldots, \tilde{\boldsymbol{\varsigma}}_{q,t}(\boldsymbol{\theta}), \ldots, 0\right), i = 1, 2, \ldots \tag{45}$$

in which to facilitate the solution, all random variables except X_t are set to zero.

In addition, the objective function is established based upon the marginal moments of the input random variables. The statistical moments of the performance function are thus calculated more accurately when the discrepancy between the estimated marginal moments of the input random variables and their exact values is lower.

The raw moments for the input random variables can be derived by

$$\overline{M}_{i,X_t} = \int_{-\infty}^{+\infty} X_t^i f_{X_t}(X_t)\mathrm{d}X_t, i = 1, 2, 3, 4 \tag{46}$$

The objective function can be defined in terms of the maximum relative error between the exact and approximate values of the marginal moments:

$$J(\boldsymbol{\theta}) = e_{\max}(\boldsymbol{\theta}) = \max\left\{\max_{1\le t\le n}\left[e_{M_{1,t}}(\boldsymbol{\theta}), e_{M_{2,t}}(\boldsymbol{\theta}), e_{M_{3,t}}(\boldsymbol{\theta}), e_{M_{4,t}}(\boldsymbol{\theta})\right]\right\} \tag{47}$$

where

$$\begin{cases} e_{M_{1,t}}(\boldsymbol{\theta}) = \dfrac{\left|M_{1,X_t} - \overline{M}_{1,X_t}\right|}{\left|\overline{M}_{1,X_t}\right|}, e_{M_{2,t}}(\boldsymbol{\theta}) = \dfrac{\left|M_{2,X_t} - \overline{M}_{2,X_t}\right|}{\left|\overline{M}_{2,X_t}\right|} \\ e_{M_{3,t}}(\boldsymbol{\theta}) = \dfrac{\left|M_{3,X_t} - \overline{M}_{3,X_t}\right|}{\left|\overline{M}_{3,X_t}\right|}, e_{M_{4,t}}(\boldsymbol{\theta}) = \dfrac{\left|M_{4,X_t} - \overline{M}_{4,X_t}\right|}{\left|\overline{M}_{4,X_t}\right|} \end{cases} \tag{48}$$

Thus, the optimal rotational angle vector $\boldsymbol{\theta}$ is determined by solving the following optimization problem:

$$\begin{cases} \text{find} & \boldsymbol{\theta} \\ \text{objective} & \min[J(\boldsymbol{\theta})] \\ \text{s.t.} & \theta_{k,l} \in [0, 2\pi] \end{cases} \tag{49}$$

The intelligent optimization algorithm [45] is applied to explore the optimal rotational angle vector $\boldsymbol{\theta}$. R-SPGR can thus calculate statistical moments more accurately. The probability distributions can be approximated by SPA technique in terms of the raw moments.

3.2.2 Advanced Mixed-Degree Cubature Formula

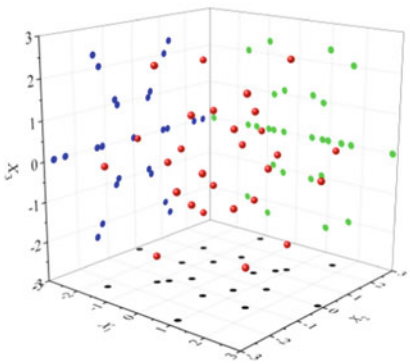
In terms of multidimensional correlations among uncertain parameters, the correlation of random variables can be quantified via the vine copula function. Drawing on the idea of rotational sparse grid method construction, the integral points of the mixed-degree cubature formula are enhanced by multidimensional transformation matrix and vine copula function to capture more sufficient information for probability distribution and correlations of random variables.

Figures 11(a–c) show the original integration points, the advanced integration points ($\boldsymbol{\theta} = [\pi/3, \pi/3, \pi/3]^{\mathrm{T}}$) and the advanced-dependence integration points, respectively. The projection numbers of the original integration points on X_1–X_2, X_1–X_3 and X_2–X_3 planes are 20, 25, and 28, respectively, while the projection numbers of the advanced integration points on different two-dimensional planes are all 28. Supposing that the dependencies among 3-dimensional random variables are represented by Clayton copula function, and Kendall's coefficients are 0.6, 0.6 and 0.78, respectively. The advanced integration points by correlation transformation are shown in Fig. 11(c), and the vine copula can be depicted the correlation among random variables.

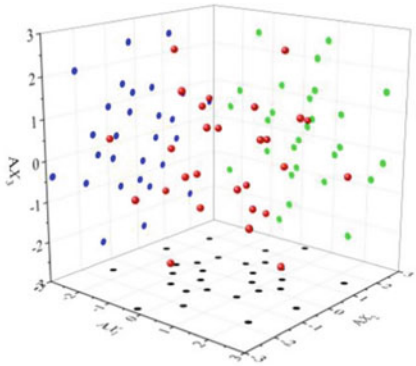
The multi-dimensional node vector of original mixed-degree cubature formula (MDCF) can be described as

$$\mathbf{U}_q = [U_{q,1}, U_{q,2}, \ldots, U_{q,n}]^{\mathrm{T}} \tag{50}$$

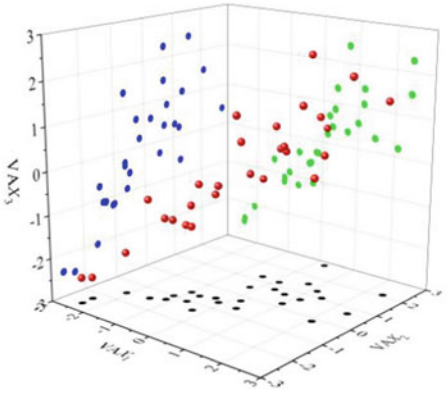
According to Eq. (41), the rotated multi-dimensional points can be obtained as



(a) Original integration points



(b) Advanced integration points



(c) Advanced-dependence integration points

Fig. 11 Integration points of mixed-degree cubature formula in 3-dimensional space

$$\tilde{\mathbf{U}}_q = \mathbf{MR}(\boldsymbol{\theta})\mathbf{U}_q = \prod_{k=1}^{n}\prod_{l=k+1}^{n}\mathbf{R}_{k,l}(\theta_{k,l})\mathbf{U}_q = [\tilde{U}_{q,1}, \tilde{U}_{q,2}, \ldots, \tilde{U}_{q,n}]^{\mathrm{T}} \tag{51}$$

Thus, the advanced multi-dimensional point matrix can be derived by

$$\tilde{\mathbf{U}} = [\tilde{\mathbf{U}}_1, \tilde{\mathbf{U}}_2, \ldots, \tilde{\mathbf{U}}_q, \ldots, \tilde{\mathbf{U}}_{N_C}]^{\mathrm{T}} \tag{52}$$

The statistical moments can be computed by the advanced mixed-degree cubature formula (AMDCF, 5th and 7th degree) [26, 46] as follows,

$$E\left[G^i(\mathbf{X})\right] = \int_{\Omega_{\mathbf{U}}} T^i(\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\mathrm{d}\mathbf{u} \approx (\pi)^{-\frac{n}{2}} \sum_{q=1}^{N_C} w_q p^i(\sqrt{2}\tilde{\mathbf{U}}_q)$$

$$
\begin{aligned}
&= (\pi)^{-\frac{n}{2}} \sum_{q=1}^{N_C} w_q \, p^i(\sqrt{2}\mathbf{MR}(\boldsymbol{\theta})\mathbf{U}_q) \\
&= (\pi)^{-\frac{n}{2}} \omega_{R,1} \\
&\left\{ \begin{array}{l} \omega_{S,1} \sum_{g=1}^{n+1} [p^i(\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_1 \mathbf{v}_g) + p^i(-\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_1 \mathbf{v}_g)] \\ +\omega_{S,2} \sum_{g=1}^{n(n+1)/2} [p^i(\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_1 \mathbf{m}_g) + p^i(-\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_1 \mathbf{m}_g)] \end{array} \right\} \\
&\quad + (\pi)^{-\frac{n}{2}} \omega_{R,2} \\
&\left\{ \begin{array}{l} \omega_{S,1} \sum_{g=1}^{n+1} [p^i(\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_2 \mathbf{v}_g) + p^i(-\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_2 \mathbf{v}_g)] \\ +\omega_{S,2} \sum_{g=1}^{n(n+1)/2} [p^i(\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_2 \mathbf{m}_g) + p^i(-\sqrt{2}\mathbf{MR}(\boldsymbol{\theta}) r_2 \mathbf{m}_g)] \end{array} \right\} \\
&= \sum_{q=1}^{N_C} \Lambda_q \, p^i(\hat{\mathbf{U}}_q)
\end{aligned} \tag{53}
$$

The integration points are given as [46]

$$
\hat{\mathbf{U}}_{q_1} = \left\{ \begin{array}{l}
\frac{\sqrt{2}}{2}\sqrt{4+2n+2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{v}_{q_1}, \\
\quad q_1 = 1, \ldots, n+1 \\
-\frac{\sqrt{2}}{2}\sqrt{4+2n+2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{v}_{(q_1-(n+1))}, \\
\quad q_1 = n+2, \ldots, 2n+3 \\
\frac{\sqrt{2}}{2}\sqrt{4+2n+2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{m}_{(q_1-(2n+2)+1)}, \\
\quad q_1 = 2n+4, \ldots, 2n+4+(\frac{n(n+1)}{2})-1 \\
-\frac{\sqrt{2}}{2}\sqrt{4+2n+2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{m}_{(q_1-2n-2-(\frac{n(n+1)}{2})+1)}, \\
\quad q_1 = 2n+4+(\frac{n(n+1)}{2})-1, \ldots, n^2+3n+2
\end{array} \right. \tag{54}
$$

$$\hat{\mathbf{U}}_{q_1+q_2} = \begin{cases} \dfrac{\sqrt{2}}{2}\sqrt{4+2n-2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{v}_{q_1}, \\ q_1 = 1, \ldots, n+1 \\ -\dfrac{\sqrt{2}}{2}\sqrt{4+2n-2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{v}_{(q_1-(n+1))}, \\ q_1 = n+2, \ldots, 2n+3 \\ \dfrac{\sqrt{2}}{2}\sqrt{4+2n-2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{m}_{(q_1-(2n+2)+1)}, \\ q_1 = 2n+4, \ldots, 2n+4+(\dfrac{n(n+1)}{2})-1 \\ -\dfrac{\sqrt{2}}{2}\sqrt{4+2n-2\sqrt{4+2n}}\,\mathbf{MR}(\boldsymbol{\theta})\mathbf{m}_{(q_1-2n-2-(\frac{n(n+1)}{2})+1)}, \\ q_1 = 2n+4+(\dfrac{n(n+1)}{2})-1, \ldots, n^2+3n+2 \end{cases} \tag{55}$$

where $q_2 = n^2+3n+2$. The corresponding weights are seen in [46, 47]. In addition, the total number of function calls is as low as $N_C = 2n^2+6n+4$.

The marginal moments can be obtained by integration [25]:

$$\mathbf{M}^t = [M_1^t, M_2^t, M_3^t, M_4^t],\ t = 1, 2, \ldots, n \tag{56}$$

The marginal moments can be also explicitly derived by AMDCF as follows,

$$\begin{aligned} \tilde{M}_1^t(\boldsymbol{\theta}) &= \sum_{q=1}^{N_C} \Lambda_q\, p(R(\mu_1), R(\mu_2), \ldots, \hat{\mathbf{U}}_{q,t}(\boldsymbol{\theta}), \ldots, R(\mu_n)) \\ &= \sum_{q=1}^{N_C} \Lambda_q\, p\Big(R(\boldsymbol{\mu}_{n-t}), \hat{\mathbf{U}}_{q,t}(\boldsymbol{\theta})\Big) \end{aligned} \tag{57}$$

$$\tilde{M}_2^t(\boldsymbol{\theta}) = \sqrt{\sum_{q=1}^{N_C} \Lambda_q \Big(p\Big(R(\boldsymbol{\mu}_{n-t}), \hat{\mathbf{U}}_{q,t}(\boldsymbol{\theta})\Big) - \tilde{M}_1^t(\boldsymbol{\theta})\Big)^2} \tag{58}$$

$$\tilde{M}_3^t(\boldsymbol{\theta}) = \sum_{q=1}^{N_C} \Lambda_q \Big(p\Big(R(\boldsymbol{\mu}_{n-t}), \hat{\mathbf{U}}_{q,t}(\boldsymbol{\theta})\Big) - \tilde{M}_1^t(\boldsymbol{\theta})\Big)^3 \Big/ \Big(\tilde{M}_2^t(\boldsymbol{\theta})\Big)^3 \tag{59}$$

$$\tilde{M}_4^t(\boldsymbol{\theta}) = \sum_{q=1}^{N_C} \Lambda_q \Big(p\Big(R(\boldsymbol{\mu}_{n-t}), \hat{\mathbf{U}}_{q,t}(\boldsymbol{\theta})\Big) - \tilde{M}_1^t(\boldsymbol{\theta})\Big)^4 \Big/ \Big(\tilde{M}_2^t(\boldsymbol{\theta})\Big)^4 \tag{60}$$

where all random variables except for the t-th variable are set to the means.

Setting $\tilde{\mathbf{M}}^t = [\tilde{M}_1^t(\boldsymbol{\theta}), \tilde{M}_2^t(\boldsymbol{\theta}), \tilde{M}_3^t(\boldsymbol{\theta}), \tilde{M}_4^t(\boldsymbol{\theta})]$, the maximum relative error of the marginal moment is derived by

$$\varepsilon^t(\boldsymbol{\theta}) = \max\left\{\left[\varepsilon_1^t(\boldsymbol{\theta}), \varepsilon_2^t(\boldsymbol{\theta}), \varepsilon_3^t(\boldsymbol{\theta}), \varepsilon_4^t(\boldsymbol{\theta})\right]\right\} = \max\left\{\left|\frac{\tilde{\mathbf{M}}^t - \mathbf{M}^t}{\mathbf{M}^t}\right|\right\} \tag{61}$$

For n-dimensional random variable, the maximum relative error can be formulated as

$$\varepsilon_{\max}(\boldsymbol{\theta}) = \max\left\{\left[\varepsilon^1(\boldsymbol{\theta}), \varepsilon^2(\boldsymbol{\theta}), \ldots, \varepsilon^n(\boldsymbol{\theta})\right]\right\} \tag{62}$$

The optimal rotational angle can be approached by minimizing $\varepsilon_{\max}(\boldsymbol{\theta})$ as follows,

$$\boldsymbol{\theta}^* = \arg\min[\varepsilon_{\max}(\boldsymbol{\theta})] \tag{63}$$

In this way, more information of probability and correlations for random variables can be obtained and AMDCF can thus calculate the statistical moments more accurately. The probability distributions are approached by Hermite polynomial model [48, 49] in case of inter-dependent random variables.

3.3 Demonstrative Dynamic Example

As shown in Fig. 12, an industrial robot dynamics problem is engaged to verify the applicability of R-SPGR and AMDCF. MCS again serves as caliber to benchmark. Fourier series and dynamic parameters can be referred to [25]. The relevant statistical data are shown in Table 7 [46]. For simplicity of demonstration, only the first joint torque accuracy is considered in this example.

Figure 13 presents the first-four order raw moments for the time interval [0, 20 s] by R-SPGR, SPGR (108 runs) and MCS (10^6 runs). It can be found that R-SPGR exhibits higher accuracy in evaluating the raw moments. It can be observed that the overall fluctuation of error with regard to R-SPGR is remarkably lower than that of SPGR method. It demonstrates that R-SPGR ensues a highly computational accuracy, thus superiority.

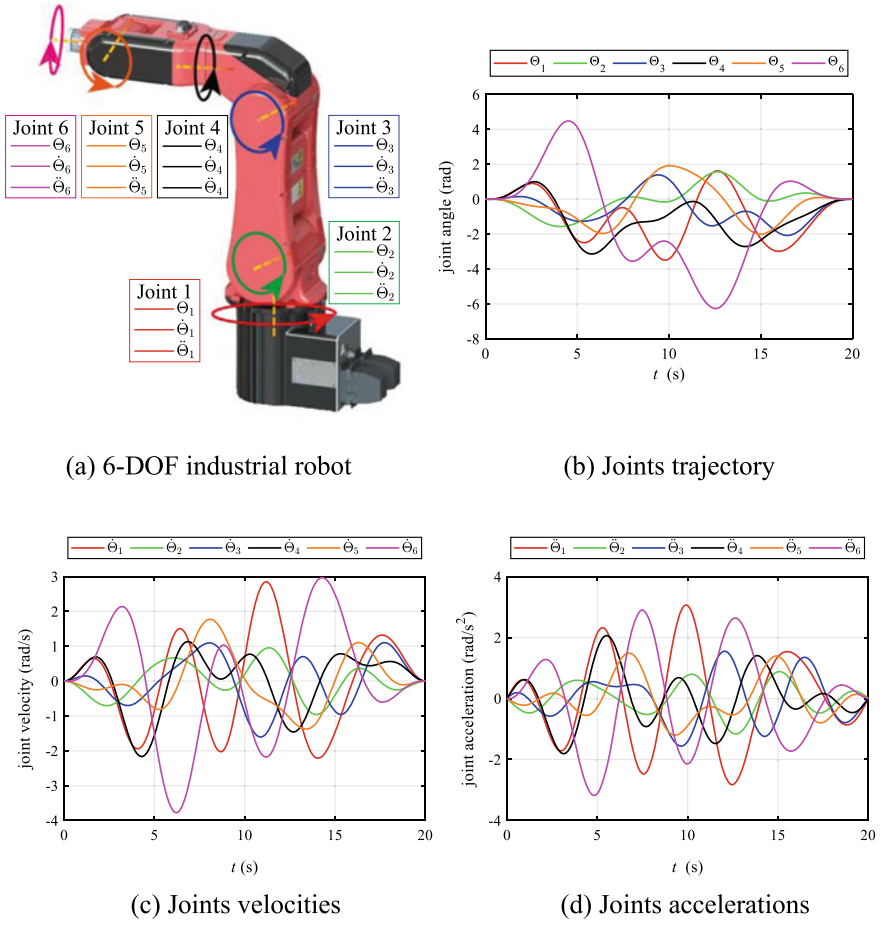
The error of the first joint torque can be computed by [25]

$$\tau_\Delta(t) = \tau_1^{\text{actual}}(t) - \tau_1^{\text{ideal}}(t)$$

where $\tau_1^{\text{actual}}(t)$ and $\tau_1^{\text{ideal}}(t)$ represent the actual and ideal torques at time instant t.

The probability distribution curves by R-SPGR, SPGR and MCS at the concerned time instants ($t = 6$, 10 s) are compared in Fig. 13. The comparison indicates that the error curve by R-SPGR fluctuates minimally. It demonstrates again the proficiency of the currently proposed R-SPGR for this dynamics example.

The correlations among the random variables in robotic dynamics could be explicit. For inter-dependent random variables, as shown in Table 8, their correlations are assumed to be vine copula. The following shows the performance of



(a) 6-DOF industrial robot

(b) Joints trajectory

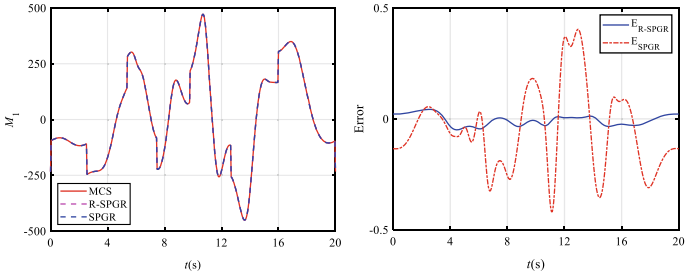
(c) Joints velocities

(d) Joints accelerations

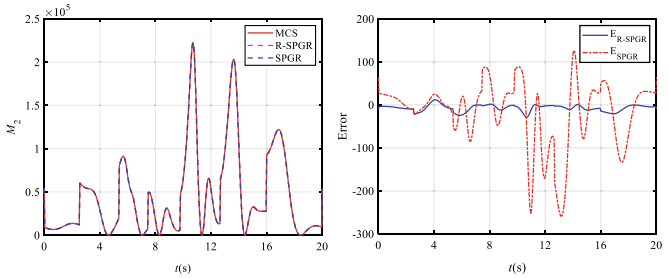
Fig. 12 Industrial robot and joint excitations [46]

Table 7 Stats for random variables in dynamic example [25]

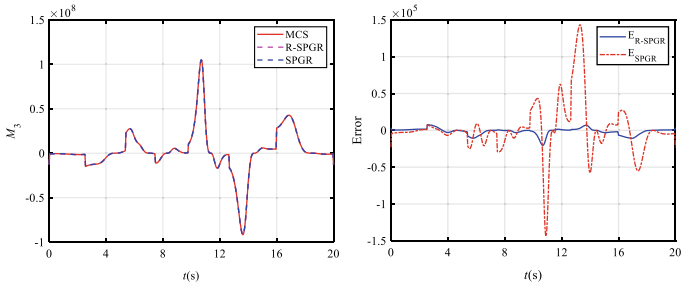
Variables	Distribution	Mean	Coefficient of variation
I_{my1}	Lognormal	4.02	0.1
I_{my2}	Lognormal	3.21	0.1
f_{c3}	Weibull	15.27	0.1
f_{c4}	Weibull	10.00	0.1
I_{xz5}	Lognormal	0.97	0.1
I_{xz6}	Lognormal	0.76	0.1



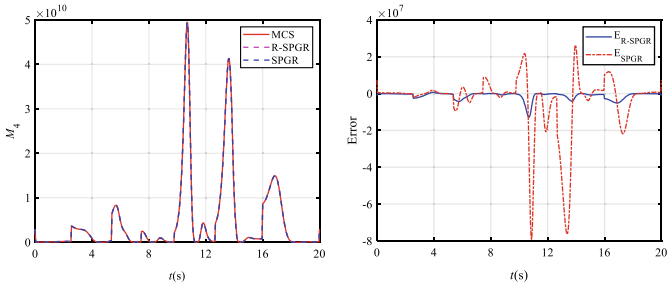
(a) Comparison of the first order raw moments



(b) Comparison of the second order raw moments



(c) Comparison of the third order raw moments



(d) Comparison of the fourth order raw moments

Fig. 13 Statistical moments in time interval [0, 20 s] [25]

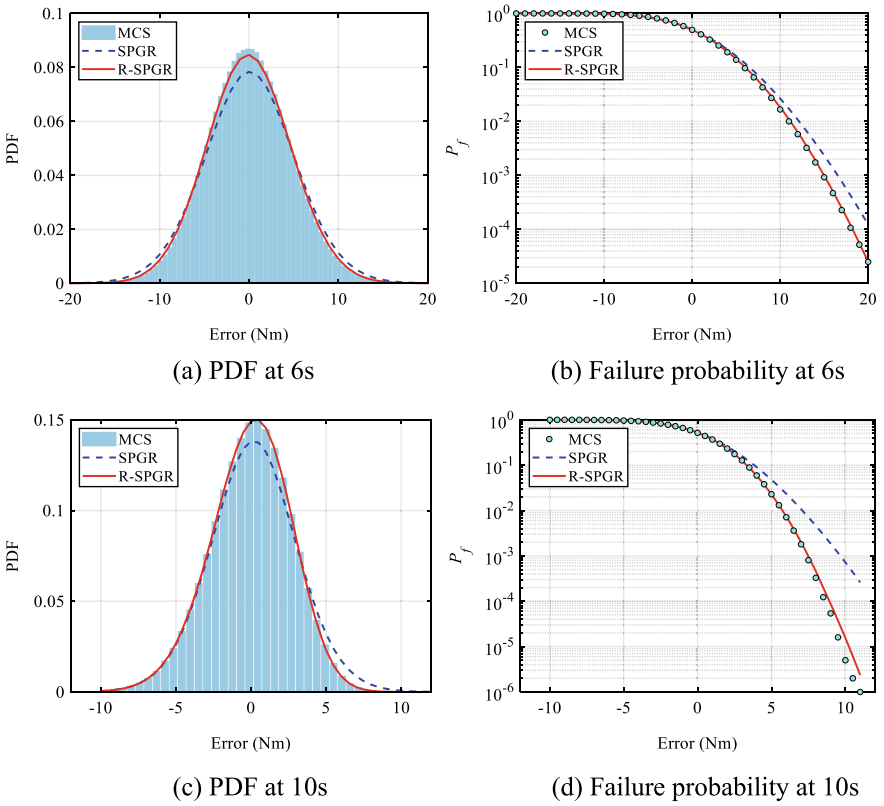


Fig. 14 PDF and failure probability subject to independent random variables [25]

AMDCF relative to dependent random variables. In addition, for dependent random variables, set $\omega_f = 0.5\pi$.

The statistical moments by the three different methods are depicted in Fig. 15. Compared with AMDCF, the error fluctuation by MDCF is larger. It proves that AMDCF can robustly attend the correlation of random variable.

Figure 16 shows PDF and failure probability by MCS, MDCF and AMDCF at the concerned time instants ($t = 12$, 14 s). It can be found that the curves by MDCF deviate more from those by MCS than the AMDCF counterpart. It can thus be stated that, for such a complex practical engineering problem, the proposed AMDCF exhibits high adaptability as well as superiority in computational accuracy.

Table 8 Pair-copulas data in D-vine copula for Example 3.3 [46]

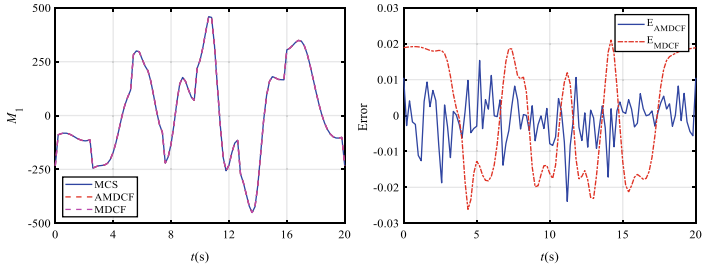
Pair-copulas	Copula type	Kendall's τ
$C\,(I_{my1}, I_{my2})$	Clayton	0.5
$C\,(I_{my2}, f_{c3})$	Clayton	0.5
$C\,(f_{c3}, f_{c4})$	Clayton	0.5
$C\,(f_{c4}, I_{xz5})$	Clayton	0.5
$C\,(I_{xz5}, I_{xz6})$	Clayton	0.5
$C\,(I_{my1}, f_{c3}\mid I_{my2})$	Frank	0.4
$C\,(I_{my2}, f_{c4}\mid f_{c3})$	Frank	0.4
$C\,(f_{c3}, I_{xz5}\mid f_{c4})$	FGM	0.09
$C\,(f_{c4}, I_{xz6}\mid I_{xz5})$	FGM	0.09
$C\,(I_{my1}, f_{c4}\mid I_{my2}, f_{c3})$	Clayton	0.5
$C\,(I_{my2}, I_{xz5}\mid f_{c3}, f_{c4})$	AMH	0.5
$C\,(f_{c3}, I_{xz6}\mid f_{c4}, I_{xz5})$	Gumbel	0.8
$C\,(I_{my1}, I_{xz5}\mid I_{my2}, f_{c3}, f_{c4})$	Gaussian	0.4
$C\,(I_{my2}, I_{xz6}\mid f_{c3}, f_{c4}, I_{xz5})$	AMH	0.5
$C\,(I_{my1}, I_{xz6}\mid I_{my2}, f_{c3}, f_{c4}, I_{xz5})$	Frank	0.5

4 Conclusions

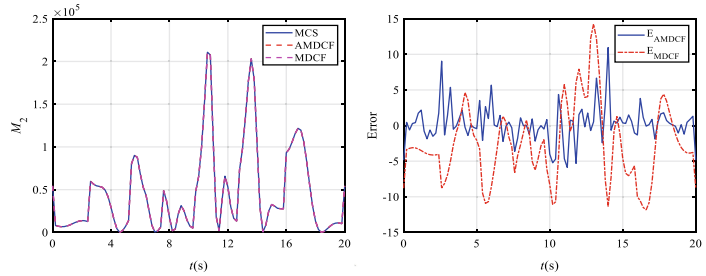
This study develops novel motion accuracy reliability analysis methods for industrial robots. The computation framework, which integrates sparse grid method, saddlepoint approximation technique, extreme value distribution theory and copula functions, is proposed to evaluate comprehensively kinematic accuracy reliability. Two efficient reliability analysis methods for dynamic accuracy, namely rotational sparse grid method and advanced mixed-degree cubature formula, are established for assessing the joint torque accuracy reliability of industrial robots subject to independent and inter-dependent uncertain parameters, respectively. The proposed methods exhibit distinct superiority compared with some other prevailing methods. The conclusions could be summarized as follows:

(1) The proposed kinematic accuracy method can comprehensively analyze the reliability of single coordinate, single point, multi-point and trajectory accuracies. The eigen-decomposition method and the copula functions effectively comprehend motion error dependencies, respectively. Compared with ME-FM and MCS methods, the proposed methods yield remarkably more accurate calculation results in the tail of the failure probability.

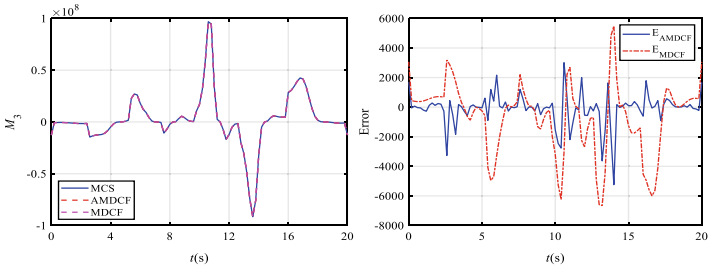
(2) For dynamic accuracy reliability, the marginal moment-matching technique and vine copula functions effectively enhance the ability of integration nodes to capture sufficient uncertainty and correlation stats. The proposed two methods are viable to calculate the statistical moments of joint torques with high accuracy and efficiency. Furthermore, the dynamic accuracy reliability of industrial robot is corroborated by different probability distribution fitting methods.



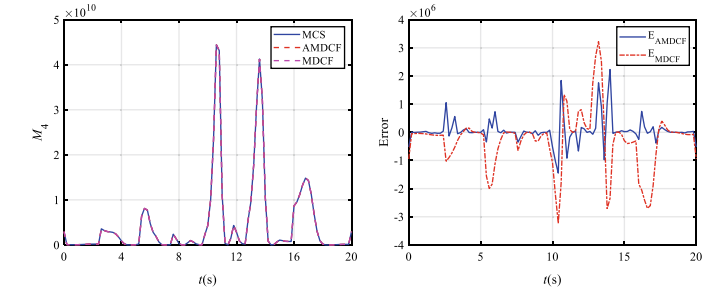
(a) Comparison of the first order raw moments



(b) Comparison of the second order raw moments



(c) Comparison of the third order raw moments



(d) Comparison of the fourth order raw moments

Fig. 15 Statistical moments subject to inter-dependent random variables in time interval [0, 20 s]

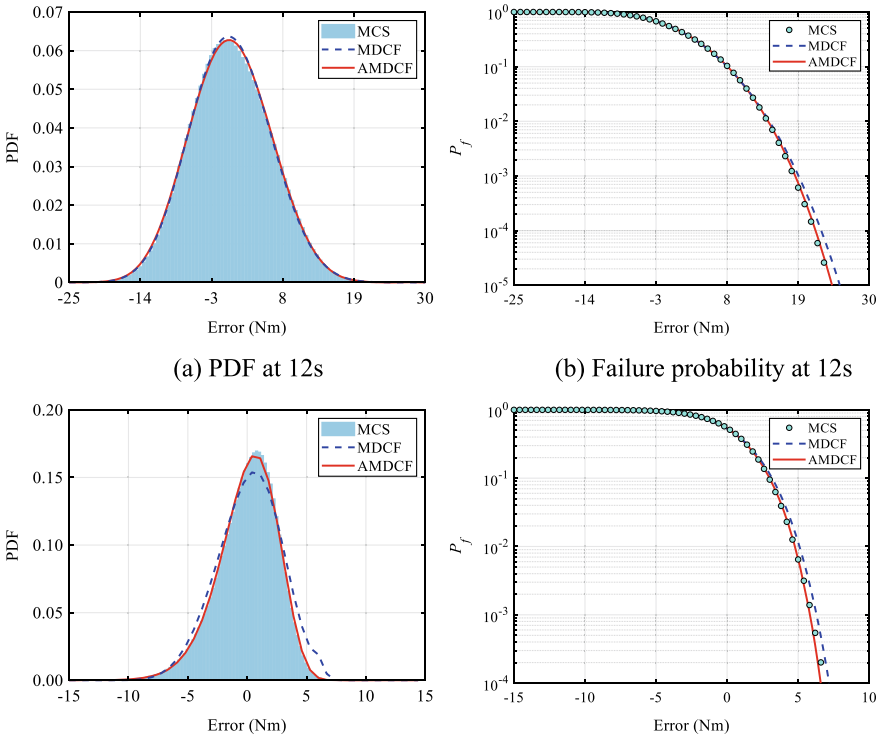


Fig. 16 PDF and failure probability subject to inter-dependent random variables [46]

(3) The kinematic and dynamic uncertain parameters influence the motion performance. This study explores transmission mechanism for the error incurred by the kinematic and dynamic uncertain parameters, accurately analyzes the mapping relationship between input and output uncertainties in industrial robots, and comprehensively evaluates the reliability of industrial robot's motion accuracy.

The proposed methods open new avenue for motion accuracy reliability assessment of industrial robot. It can serve as theoretical support for industrial robot reliability-based design. A note of attention is that this study focuses on the effect of stochastic uncertainty parameters on the motion performance of industrial robots. However, cognitive uncertainties such as actual operating actions and ambient temperature changes may also negate the accuracy and reliability. Therefore, our future study will attend the reliability analysis and optimal design of industrial robots subject to multi-source uncertainties.

Acknowledgements The authors would like to acknowledge the financial supports from the National Natural Science Foundation of China (Grant Nos. 52275244 and 51905146), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 51621004).

References

1. Hafezipour M, Khodaygan S (2017) An uncertainty analysis method for error reduction in end-effector of spatial robots with joint clearances and link dimension deviations. Int J Comput Integr Manuf 30(6):653–663
2. Zhang D, Han X (2020) Kinematic reliability analysis of robotic manipulator. J Mechan Des 142(4)
3. Zhao Q, Hong J (2022) An analytical framework for local and global system kinematic reliability sensitivity of robotic manipulators. Appl Math Model 102:331–350
4. Chen J, Chen L, Qian L, Chen G, Zhou S (2022) Time-dependent kinematic reliability analysis of gear mechanism based on sequential decoupling strategy and saddle-point approximation. Reliab Eng Syst Saf 220:108292
5. Lee U, Lee I (2022) Efficient sampling-based inverse reliability analysis combining Monte Carlo simulation (MCS) and feedforward neural network (FNN). Struct Multidiscip Optim 65(1):1–22
6. Keshtegar B, Seghier MEAB, Zio E, Correia JA, Zhu SP, Trung NT (2021) Novel efficient method for structural reliability analysis using hybrid nonlinear conjugate map-based support vector regression. Comput Methods Appl Mech Eng 381:113818
7. Meng Z, Guo L, Wang X (2022) A general fidelity transformation framework for reliability-based design optimization with arbitrary precision. Struct Multidiscip Optim 65(1):1–16
8. Hu Z, Du X (2019) Efficient reliability-based design with second order approximations. Eng Optim 51(1):101–119
9. Dudzik A, Potrzeszcz-Sut B (2021) Hybrid approach to the first order reliability method in the reliability analysis of a spatial structure. Appl Sci 11(2):648
10. Yi J, Wu F, Zhou Q, Cheng Y, Ling H, Liu J (2021) An active-learning method based on multi-fidelity Kriging model for structural reliability analysis. Struct Multidiscip Optim 63(1):173–195
11. Yang M, Zhang D, Wang F, Han X (2022) Efficient local adaptive Kriging approximation method with single-loop strategy for reliability-based design optimization. Comput Methods Appl Mech Eng 390:114462
12. Ameryan A, Ghalehnovi M, Rashki M (2022) AK-SESC: a novel reliability procedure based on the integration of active learning kriging and sequential space conversion method. Reliab Eng Syst Saf 217:108036
13. Xu J, Du Y, Zhou L (2021) A multi-fidelity integration rule for statistical moments and failure probability evaluations. Struct Multidiscip Optim 64(3):1305–1326
14. Fan W, Liu R, Ang AH, Li Z (2018) A new point estimation method for statistical moments based on dimension-reduction method and direct numerical integration. Appl Math Model 62:664–679
15. Rao SS, Bhatti PK (2001) Probabilistic approach to manipulator kinematics and dynamics. Reliab Eng Syst Saf 72(1):47–58
16. Kim J, Song WJ, Kang BS (2010) Stochastic approach to kinematic reliability of open-loop mechanism with dimensional tolerance. Appl Math Model 34(5):1225–1237
17. Pandey MD, Zhang X (2012) System reliability analysis of the robotic manipulator with random joint clearances. Mech Mach Theory 58:137–152

18. Wu J, Zhang D, Liu J, Han X (2019) A moment approach to positioning accuracy reliability analysis for industrial robots. IEEE Trans Reliab 69(2):699–714
19. Lara-Molina FA, Dumur D (2021) A fuzzy approach for the kinematic reliability assessment of robotic manipulators. Robotica 39(12):2095–2109
20. Zhao Q, Guo J, Zhao D, Yu D, Hong J (2021) Time-dependent system kinematic reliability analysis for robotic manipulators. J Mechan Des 143(4)
21. Zhang D, Shen S, Wu J, Wang F, Han X (2023) Kinematic trajectory accuracy reliability analysis for industrial robots considering intercorrelations among multi-point positioning errors. Reliab Eng Syst Saf 229:108808
22. Yang B, Yang W (2023) Modular approach to kinematic reliability analysis of industrial robots. Reliab Eng Syst Saf 229:108841
23. Cao L, Liu J, Zhang J, Jiang C, Zhang D (2023) Positioning accuracy reliability analysis of industrial robots considering epistemic uncertainty and correlation. J Mech Des 145(2):023303
24. Chen H, Wang R, Dong Y, He S (2022) Dynamic reliability analysis for configuration transformation of a controllable metamorphic palletizing robot. Proc Inst Mech Eng C J Mech Eng Sci 236(10):5208–5222
25. Wu J, Zhang D, Jiang C, Han X, Li Q (2021) On reliability analysis method through rotational sparse grid nodes. Mech Syst Signal Process 147:107106
26. Xu J, Zhang Y, Dang C (2020) A novel hybrid cubature formula with Pearson system for efficient moment-based uncertainty propagation analysis. Mech Syst Signal Process 140:106661
27. Rahman S, Xu H (2004) A univariate dimension-reduction method for multi-dimensional integration in stochastic mechanics. Probab Eng Mech 19(4):393–408
28. Xiong F, Xiong Y, Greene S, Chen W, Yang S (2009, January) A new sparse grid based method for uncertainty propagation. In: International design engineering technical conferences and computers and information in engineering conference (vol 49026, pp 1205–1215)
29. Wu J, Zhang D, Liu J, Jia X, Han X (2020) A computational framework of kinematic accuracy reliability analysis for industrial robots. Appl Math Model 82:189–216
30. Liu J, Hu Y, Xu C, Jiang C, Han X (2016) Probability assessments of identified parameters for stochastic structures using point estimation method. Reliab Eng Syst Saf 156:51–58
31. Saberzadeh Z, Razmkhah M (2022) Reliability of degrading complex systems with two dependent components per element. Reliab Eng Syst Saf 222:108398
32. Lu H, Zhu Z, Zhang Y (2019) A hybrid approach for reliability-based robust design optimization of structural systems with dependent failure modes. Eng Optim
33. Wang Z, Zhang X, Huang HZ, Mourelatos ZP (2016) A simulation method to estimate two types of time-varying failure rate of dynamic systems. J Mechan Des 138(12)
34. Okoro A, Khan F, Ahmed S (2023) Dependency effect on the reliability-based design optimization of complex offshore structure. Reliab Eng Syst Saf 231:109026
35. Love PE, Tenekedjiev K (2022) Understanding near-miss count data on construction sites using greedy D-vine copula marginal regression: a comment. Reliab Eng Syst Saf 217:108021
36. Li X, Zhang W, He L (2020) Bayes theorem–based and copula-based estimation for failure probability function. Struct Multidiscip Optim 62(1):131–145
37. Jiang C, Zhang W, Han X, Ni BY, Song LJ (2015) A vine-copula-based reliability analysis method for structures with multidimensional correlation. J Mech Des 137(6):061405
38. Jia XY, Jiang C, Fu CM, Ni BY, Wang CS, Ping MH (2019) Uncertainty propagation analysis by an extended sparse grid technique. Front Mech Eng 14(1):33–46
39. Huang BQ, Du XP (2006) Uncertainty analysis by dimension reduction integration and saddlepoint aporoximations. J Mech Des 128(1):26–33
40. Hu Z, Du X (2018) Saddlepoint approximation reliability method for quadratic functions in normal variables. Struct Saf 71:24–32
41. Zhu LS, Lu H, Zhang YM (2021) A system reliability estimation method by the fourth moment saddle point approximation and copula functions. Qual Reliab Eng Int 37(6):2950–2969
42. Wu H, Du X (2020) System reliability analysis with second-order saddlepoint approximation. ASCE-ASME J Risk Uncertainty Eng Syst Part B Mechan Eng 6(4):041001

43. Jin J, Gans N (2015) Parameter identification for industrial robots with a fast and robust trajectory design approach. Robot Comput-Integr Manuf 31:21–29
44. Xu J, Kong F (2018) An efficient method for statistical moments and reliability assessment of structures. Struct Multidiscip Optim 58(5):2019–2035
45. Pozna C, Precup RE, Horváth E, Petriu EM (2022) Hybrid particle filter-particle swarm optimization algorithm and application to fuzzy controlled servo systems. IEEE Trans Fuzzy Syst 30(10):4286–4297
46. Zhang D, Shen S, Jiang C, Han X, Li Q (2022) An advanced mixed-degree cubature formula for reliability analysis. Comput Methods Appl Mech Eng 400:115521
47. Meng D, Miao L, Shao H, Shen J (2018) A seventh-degree cubature Kalman filter. Asian J Control 20(1):250–262
48. Xu J, Wu Z, Lu ZH (2022) An adaptive polynomial skewed-normal transformation model for distribution reconstruction and reliability evaluation with rare events. Mech Syst Signal Process 169:108589
49. Zhao Z, Lu ZH, Zhao YG (2022) Time-variant reliability analysis using moment-based equivalent Gaussian process and importance sampling. Struct Multidiscip Optim 65(2):1–17

Reliability of Wireless Body Area Networks

Liudong Xing, Guilin Zhao, and Qun Zhang

Abstract A wireless body area network (WBAN) is a network of low-power devices including smart sensors situated in, on, or around the human body to monitor the physiological and motion information for healthcare, military, sports, security, fire-fighting, as well as other applications and purposes. Reliability is one of the major changes to address for delivering the desired quality of services of WBANs. In this chapter, a critical review of WBAN reliability-related literature is conducted, covering reliability modeling, analysis, and designs. A reliability model is also presented for WBANs subject to the probabilistic function dependence and associated probabilistic isolation and competing behaviors. The model is demonstrated through a case study on the reliability analysis of a WBAN patient monitoring system.

Keywords Competition · Probabilistic function dependence (PFD) · Probabilistic isolation · Reliability · Wireless body area network (WBAN)

1 Introduction

A wireless body area network (WBAN), also known as a body sensor network (BSN), is a wireless network of wearable and low-power computing devices, including smart sensors situated in, on, or around the human body to monitor the physiological and

L. Xing (✉)
Department of Electrical and Computer Engineering, University of Massachusetts, Dartmouth, MA, USA
e-mail: lxing@umassd.edu

G. Zhao
School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, Sichuan, China
e-mail: guilinzhao@swjtu.edu.cn

Q. Zhang
Department of Electrical and Computer Engineering and Technology, Minnesota State University, Mankato, MN, USA
e-mail: qun.zhang@mnsu.edu

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_4

motion information [1, 2]. The WBAN technology started around 1995 with the idea of collecting personal data via portable devices and implementing wireless communication technologies on personal area networks. In the past decade, due to the rapid development of the Internet and mobile computing technologies, WBAN has been significantly developed as one of the most promising fields and has been applied in medical applications (i.e., healthcare) and non-medical domains (such as military, sports, security, firefighting, etc.). For example, using healthcare WBANs, health professionals can keep records of chronic pulmonary patients to design scientific patient management to relieve their health conditions [3]. In the military application of WBANs, diverse types of sensors (e.g., acoustic, seismic, radar, optical and magnetic) as well as global positioning systems (GPS) and the global system for mobile communication (GSM) modem, operate collaboratively for the collection and processing of an enormous amount of real-time information to protect soldiers' safety and improve the combat capability of an individual soldier (or a small unit) in the battlefield [4–6].

The reliability issue is one of the most critical challenges for the WBAN to deliver the desired quality of service (QoS) [7–9]. Consider the above mentioned chronic pulmonary patients care example. If the patient's respiratory failure is not detected or the WBAN system transmits erroneous data to caregivers, their safety may be threatened when they need emergency handling from the care provider. Likewise, for the military equipment (e.g., the Tactical Assault Light Operator Suit used for detecting fatigue of special forces soldiers in unsafe environments), the malfunction of hardware, high network latency due to multiple routers, or unnecessary bandwidth waste due to poorly optimized resource allocation may put soldiers into a dangerous situation. Due to the life-critical nature of the WBAN applications, reliability modeling, analysis, and designs are imperative for the robust operation of WBAN systems.

Some recent review articles offer comprehensive discussions that address the current communication technologies for WBANs (e.g., Bluetooth, ZigBee, Wi-Fi, IEEE 802.15.6, etc.) and the possible incorporation of new technologies into WBANs (e.g., Software Defined Networking, Energy Harvesting, and Blockchain technology) [7, 10, 11]. Different from these existing works, in this chapter, we conduct a critical review of WBAN reliability-related literature, covering reliability modeling, analysis, and designs. A specific reliability model is also presented for WBANs subject to the probabilistic function dependence (PFD) and associated probabilistic isolation and competing behaviors.

The remainder of this chapter is organized as follows: Section 2 explains the major components that constitute a WBAN system and the dependent relationship among the WBAN components. Section 3 presents the literature review on the WBAN reliability-related works. Section 4 presents the WBAN reliability analysis method addressing the PFD and related isolation and competing behaviors. Section 5 presents the case study demonstrating the reliability model. Section 6 concludes the chapter.

2 WBAN Components and Dependent Relationship

A WBAN is a wireless network of low-power devices (smart sensors and actuators), which may be implanted into the human body or surface-mounted on the body or nearby the body by carrying them in bags or clothes pockets [12, 13]. Because the energy consumed by data transmissions over wireless links is inversely proportional to the square of the transmission distance [14], relays are typically used for conserving the limited energy of battery-powered biosensors.

The WBAN also contains a local processing unit (LPU) or a sink device, which can be a smart phone. The LPU may conduct some simple processing on the data collected. In the case of detecting any abnormalities, an immediate alert is issued to the person wearing the system. The LPU also functions as a router between the WBAN components and a central server using the mobile/wireless networks and the Internet. After receiving the data, the central server feeds the data into its database and performs analysis of the data. Dependent on the degree of abnormalities detected from the data analysis, the server may interact with different parties. For example, for WBANs used in the healthcare domain, based on the abnormalities detected, the server may interact with the patient's family members, a local physician, or the emergency unit of a nearby hospital.

Figure 1 illustrates the major components and general structure of a WBAN system.

There exists PFD between a relay and biosensors that use the relay for transmitting their sensed information to the sink device. In the case of the relay being operating correctly, the biosensors send the data to the sink through the relay node for conserving their limited battery power. In the case of the relay failing, if the biosensor has adequate battery power to enable a direct communication with the sink, the data can be transmitted; otherwise, the biosensor is isolated and thus its sensed data cannot reach the sink. So, the PFD incurs the probabilistic isolation of the biosensors when the relay malfunctions. Moreover, the PFD can incur competitions between the failure of the relay and propagated failures (PFs) of dependent biosensors in the time domain. For example, biosensors may be subject to PFs due to jamming attacks (launched by continually sending interference data from a compromised biosensor to block legitimate data transmissions to the sink [15, 16]). If such PFs take place before the relay fails, the propagation effect may cause extensive damages and even crash the entire WBAN. However, if the relay fails before the PFs, these biosensors and thus their PFs are isolated with certain probabilities, dependent on their remaining battery power levels. In this case, the WBAN may continue to operate, dependent on the remaining redundancy. In Sect. 4, a combinatorial reliability model is presented to address the PFD and associated probabilistic isolation and competition behaviors in the WBAN reliability analysis.

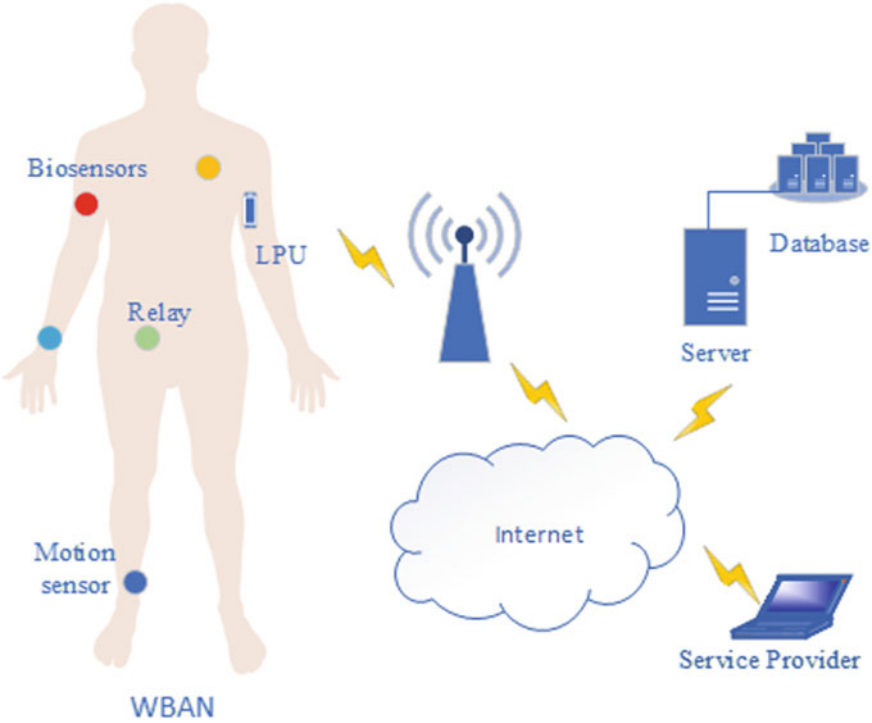


Fig. 1 Structure of a typical WBAN system

3 Related Work on WBAN Reliability

In the past decade, considerable research efforts have been devoted to the reliability modeling, analysis, and designs of WBAN systems. This section presents a critical review of these efforts.

3.1 Reliability Modeling and Analysis of WBANs

A number of proposed methods on reliability modeling and analysis of WBANs are targeted towards combinatorial models. Others are geared towards coping with state-space models.

3.1.1 Combinatorial Models

The combinatorial models for WBAN reliability modeling and analysis systematically decompose the original complex reliability issue into a set of reduced reliability problems that can be efficiently solved by the Binary Decision Diagram (BDD)-based or Multi-valued Decision Diagram (MDD)-based analytical approach. The failure criteria of the original WBAN system are represented by the dynamic fault trees, where a set of static fault tree gates (e.g., AND, OR, and k-out-of-n gates) and dynamic fault tree gates (e.g., Function DEPendency and PFD gates) is connected to represent the different combinations and dynamic behaviors contributing to the WBAN failure, such as the PFD, cascading failures, phased-mission behavior, sequence dependence, warm and cold standby sparing, etc. [17].

For example, in [16], a BDD-based combinatorial approach was proposed for reliability modeling and analysis of WBAN systems with one PFD group or multiple independent and non-overlapped PFD groups. Specifically, by applying the simple and efficient algorithm (SEA) [18] to the dynamic fault tree (DFT) model of the WBAN system, the proposed method separates the PF of the relay first. Building on the failure competing events (FCEs) and the probabilistic function dependence cases (PFDCs) defined, a set of reduced system fault tree model without the probabilistic competing behavior is generated, which can then be evaluated using the SEA-based procedure and the BDD-based method. Different parameter settings were used to illustrate the interacting improvement and deterioration effects caused by the failure isolation. In [19], the analytical model of [16] was extended to address WBAN systems where the correlation is involved due to multiple biosensors sharing a relay and a biosensor being able to communicate with the sink via more than one relay. In [20], an MDD-based combinatorial approach was applied to analyze the reliability of WBAN systems with the phased-mission behavior, where systems involve night sleep and daytime activities phases in a daily cycle.

In [21], a combinatorial model was proposed to address the random failure isolation time existing in realistic WBAN systems where in the case of the relay malfunctioning, the direct communication between dependent biosensors and the sink device was established and was able to last for a random duration, depending on the power level of the biosensor's battery, the distance between the biosensor and the sink device, and the battery discharge model. In [22], a combinatorial model was suggested for the reliability modeling and analysis of WBAN systems involving correlated PFD groups and random failure propagation time for dependent biosensors undergoing PFs.

In the above-mentioned works [16, 19–22], different types of time-to-failure distributions, including exponential, Weibull, and log-normal distributions, were adopted for biosensors to illustrate that the proposed combinatorial methods are applicable to arbitrary types of lifetime distributions of WBAN components.

3.1.2 State-Space Models

State space models used for the reliability modeling and analysis of WBAN systems include the Markov chain methods and the Petri nets. For example, in [23], a Semi-Markov process was applied to model the nodes with misbehaviors in WBAN systems, and evaluated to derive the WBAN reliability in terms of the network survivability; the node states are classified into cooperative, selfish, malicious, failed, and route states. In [24], the Markov chain model was used to analyze the dynamic gate blocks, which were introduced by the proposed Forest Topology Three Tire (FTTT) framework against malicious attacks, such as the denial-of-service attacks in WBAN systems. In [25], the Markov chain model was used to analyze the failures and their influence on the availability indicator of WBAN systems, where nineteen states ranging from the "normal condition system" to the "failure of the IoT healthcare system" were defined. In [26], the Markov chain model was utilized for the reliability and mean time to failure (MTTF) estimation of WBAN systems with the consideration of failure rates of biosensors, the patient's health condition, the accuracy of anomaly detection algorithms (measured by the detection rate and false positive rate), and the accuracy of transient fault correction measures.

Some research efforts were also dedicated to the Petri net-based models for reliability modeling and analysis of WBANs. For example, in [27] a Petri Net service model was suggested to represent the data flow of sensors so as to track the data from the originating source; the reliability was preserved in the content of the data traceability because data traceability may prevent the transparent audit of the streamed medical data from possible compromises. In [28], a Stochastic Petri net (SPN) model was proposed to address the availability of e-health systems that comprise a WBAN system, internet-connected gateways, and cloud and big data support; a multi-objective optimization solution was proposed to consider the trade-off between the system availability and the cost.

3.2 Reliability Design for WBANs

The reliability design for WBAN systems includes several issues from different perspectives.

- From the perspective of the challenges of WBANs: as the components are typically powered by batteries, the power consumption is one of the major constraints while addressing reliability optimization of the WBAN system.
- From the perspective of reliability engineering: abundant works that address the reliable WBAN system designs are based on fault-tolerant mechanisms coping with possible faults to prevent them leading to the WBAN failure.

- From the perspective of WBAN enabling technologies: for example, allocating strategies for WBAN nodes, resource scheduling strategies, communication protocols designs, and artificial-intelligent (AI)-based technology are explored for designing reliable WBANs.

Considering that there are different levels in a WBAN, including node, communication, data, and system levels, the above issues related to reliability designs at each level are discussed below.

3.2.1 Node Level

The node level contains physical entities such as wearable/implantable biosensors, actuators, relays/hubs/coordinators, and LPU. A reliable WBAN can be obtained by optimizing the performance of those physical nodes. For example, in [29], the optimization was performed on parameters for an ultra-low power WBAN considering the power consumed by diverse hardware components such as the low noise amplifier, squaring device, analog-to-digital converter, and clock.

Fault tolerance through redundancy may ensure the reliable operation of WBAN nodes. For example, in [30], two sinks were placed in the WBAN system and work as backup for each other.

One of WBAN enabling technologies adopted at the node level is machine learning, including but not limited to Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Support Vector Machines (SVM), and Decision Trees (DT). For example, in [31], these machine learning technologies were applied based on the WBAN traffic dataset, the biological dataset, and the synthesized dataset to detect faults in WBAN systems. Another WBAN enabling technology at the node level is the node allocation technology, which explores the optimal number and positions of the nodes in the WBAN. For example, in [32], the best position for the center node in the WBAN system was explored considering the energy consumption and different places of the center node (at the right hip and chest).

3.2.2 Communication Level

The communication level contains software components such as application programs, Medium Access Control (MAC) protocols, routing protocols, and communication protocols (including Bluetooth, Wi-Fi, Zigbee, etc.). The reliability at this level can be improved by applying multi-objective optimization for different software components. For example, in [33], the optimization was performed for the routing protocol considering the energy efficiency of biosensors and the WBAN lifetime.

Another way to improve the reliability at the communication level is through using the fault-tolerant mechanism. For example, in [34], the fault-tolerant communication approach was designed to improve the reliability of WBANs by minimizing

the channel impairment and fading effect with the cooperative communication via multiple biosensors and the gateway.

The MAC protocols design has also been investigated to improve the WBAN reliability. For example, in [35], compared to the traditional Time Division Multiple Access (TDMA) MAC protocols adopted by the WBAN legacy standards (such as IEEE 802.15.4 and IEEE 802.15.6), two new TDMA techniques were suggested to achieve a more reliable and energy efficient WBAN system by sufficiently considering the reliability of nodes and the channel status.

3.2.3 Data Level

The data level reliability design has been performed related to either the optimization problem or the fault-tolerant mechanism. As an illustration of the optimization, in [36], the reliability of the transmission of emergency critical sensory data was optimized by a resource scheduling strategy based on the deep reinforcement learning algorithm. As an illustration of the fault-tolerant mechanism, in [37], the data transmission reliability and efficiency were enhanced by introducing the backup path for data transmissions.

3.2.4 System Level

The system level reliability design copes with the hardware and software implementation of WBAN systems. The implementation of a WBAN system involves the construction from modules to system architecture, tuning design parameters to optimize the QoS indexes, and designing experiments to verify and validate the performance of the prototype WBAN system. For example, in [38], a reliable WBAN system was designed and implemented, and the QoS indexes of packet error rate and frame error rate were evaluated. In [39], a cooperative software-hardware framework including a simulator, a hardware platform, and a code generator, was implemented for WBANs; the proposed cooperative framework can verify and compare the performance (in terms of energy consumption and throughput) of WBAN systems adopting different protocols, such as Adaptive Data Transmission MAC (AT-MAC), Timeout-MAC (T-MAC), Sensor-MAC (S-MAC), and IEEE 802.15.4 MAC protocols. In [40], the simulation and implementation of narrowband physical layer architecture for the WBAN system based on IEEE 802.15.6 MAC protocol was performed, where the Bose–Chaudhuri–Hocquenghem (BCH) block coding for error detection and correction was included to obtain the high reliability in data transmissions.

Table 1 summarizes the reliability design works reviewed above.

Table 1 Examples of WBAN reliability design methods

Reliability design	Representative references at different levels
Optimization	• Node level: [29] • Communication level: [33] • Data level: [36]
Fault tolerance	• Node level: [30] • Communication level: [34] • Data level: [37]
WBAN enabling technologies	• Node level: [31, 32] • Communication level: [35] • Data level: [36] • System level: [38–40]

4 Reliability Model Considering PFD

This section presents a combinatorial reliability analysis procedure for a WBAN system while addressing the PFD and associated probabilistic isolation and competing behaviors described in Sect. 2. The procedure includes the following five steps.

Step 1 Based on the law of total probability, separate effects of PFs of all non-dependent biosensors (NDB) from the WBAN unreliability (denoted by *UR*) solution combinatorics using

$$UR = 1 - NP + NP * CF. \tag{1}$$

NP in (1) denotes the probability that none of the NDBs experience PFs. Let q_P represent the PF probability of an NDB. Then *NP* is computed as the product of $(1 - q_P)$ for all the NDBs.

CF in (1) denotes the conditional probability that the WBAN considered becomes failed conditioned on no PFs happening to NDBs. In other words, only local failures of NDBs are considered in the evaluation of *CF*. However, the local failure (LF) probability of each NDB (denoted by q_L) must be revised to a conditional probability (denoted by q) given that this NDB does not experience any PFs. When an NDB's LF and PF take place independently, $q = q_L$; when they happen in a mutually exclusive manner, $q = q_L/(1 - q_P)$. The evaluation of *CF* is performed through the next three steps.

Step 2 Based on the relays in the WBAN, generate the combined relay event (CRE) space. Suppose there are *M* relays, denoted by $r_1, \ldots, r_M$. Then there are 2^M CREs in the space, summarized in Table 2. The occurrence probability of each CRE, denoted by $\mathrm{P}(CRE_k)$ $(k = 0, 1, \ldots, 2^M - 2, 2^M - 1)$ can be evaluated based on its definition, and the conditional failure probability q derived from the input failure parameters of the relays.

Table 2 CRE space built based on relays in the WBAN

Notation	Event definition	Event explanation
CRE_0	$\bar{r}_1 \cap \cdots \cap \bar{r}_{M-1} \cap \bar{r}_M$	None of the relays malfunction
CRE_1	$\bar{r}_1 \cap \cdots \cap \bar{r}_{M-1} \cap r_M$	Only relay r_M malfunctions
…	…	…
CRE_{2^M-2}	$r_1 \cap \cdots \cap r_{M-1} \cap \bar{r}_M$	All relays except r_M malfunction
CRE_{2^M-1}	$r_1 \cap \cdots \cap r_{M-1} \cap r_M$	All relays malfunction

Step 3 According to the law of total probability and the CRE space built in Step 2, we evaluate CF in (1) as

$$CF = \sum_{k=0}^{2^M-1} \mathrm{P}(\text{WBAN fails}|CRE_k) \times \mathrm{P}(CRE_k) = \sum_{k=0}^{2^M-1} W_k \times \mathrm{P}(CRE_k). \quad (2)$$

The evaluation method of $\mathrm{P}(\text{WBAN fails}|CRE_k) = W_k$ is discussed under the following two cases:

- Under event CRE_0, all the relays are operating correctly, thus no isolation or competition happens. In this case, to evaluate the corresponding W_0, all the events modeling the relay failures in the WBAN reliability model (for example, the fault tree) are replaced with constant "0" (meaning FALSE) and Boolean reductions are then applied to generate a reduced WBAN fault tree model. The resulting fault tree can be evaluated using any existing fault tree analysis method to obtain W_0. In this work, the BDD-based method is adopted for the fault tree analysis [41].
- Under event CRE_k ($k = 1 \ldots 2^M - 1$), at least one relay malfunctions incurring probabilistic isolation and competitions. If CRE_k happens first, then the probabilistic isolation occurs (denoted by subcase $CRE_{k,I}$); if any dependent biosensor's PF happens first, then the propagation effect occurs (denoted by subcase $CRE_{k,P}$). Based on these two subcases, Step 4 elaborates the evaluation of W_k ($k = 1 \ldots 2^M - 1$).

Step 4 Based on the set of dependent biosensors affected by CRE_k, build the isolation event (IE) space. Suppose there are n_k dependent biosensors affected by CRE_k (each denoted by $db_{k,l}$, $l = 1 \ldots n_k$), constituting set S_k (for $k = 1 \ldots 2^M - 1$). Table 3 summarizes the definitions of all the IEs defined under CRE_k.

Table 3 The IE space under CRE_k

Notation	IE definition	IE explanation
$IE_{k,0}$	$\overline{db}_{k,1} \cap \cdots \cap \overline{db}_{k,n_k-1} \cap \overline{db}_{k,n_k}$	None of the dependent biosensors are isolated
$IE_{k,1}$	$\overline{db}_{k,1} \cap \cdots \cap \overline{db}_{k,n_k-1} \cap db_{k,n_k}$	Only biosensor db_{k,n_k} is isolated
…	…	…
$IE_{k,2^{n_k}-2}$	$db_{k,1} \cap \cdots \cap db_{k,n_k-1} \cap \overline{db}_{k,n_k}$	All dependent biosensors except d_{k,n_k} are isolated
$IE_{k,2^{n_k}-1}$	$db_{k,1} \cap \cdots \cap db_{k,n_k-1} \cap db_{k,n_k}$	All n_k dependent biosensors are isolated

Based on $CRE_{k,I}$ and $CRE_{k,P}$ defined in Step 3 and the IE space defined in Table 3, $W_k \times \mathrm{P}(CRE_k) = P(\text{WBAN fails}|CRE_k) \times P(CRE_k)$ in (2) for $k = 1 \ldots 2^M - 1$ can be evaluated by applying the law of total probability as

$$\begin{aligned} & W_k \times \mathrm{P}(CRE_k) = P(\text{WBAN fails}|CRE_k) \times P(CRE_k) \\ & = \sum_{j=0}^{2^{n_k}-1} \left[\begin{array}{c} P\left(\text{WBAN fails}|IE_{k,j} \cap CRE_{k,P}\right) \times P\left(IE_{k,j} \cap CRE_{k,P}\right) + \\ P\left(\text{WBAN fails}|IE_{k,j} \cap CRE_{k,I}\right) \times P\left(IE_{k,j} \cap CRE_{k,I}\right) \end{array} \right] \\ & = \sum_{j=0}^{2^{n_k}-1} \left[\begin{array}{c} P\left(IE_{k,j} \cap CRE_{k,P}\right) + \\ P\left(\text{WBAN fails}|IE_{k,j} \cap CRE_{k,I}\right) \times P\left(IE_{k,j} \cap CRE_{k,I}\right) \end{array} \right] \end{aligned} \tag{3}$$

In (3), $\mathrm{P}(IE_{k,j} \cap CRE_{k,P})$ and $\mathrm{P}(IE_{k,j} \cap CRE_{k,I})$ can be evaluated based on the input failure parameters of WBAN components and values of isolation factors. To evaluate $P\left(\text{WBAN fails}|IE_{k,j} \cap CRE_{k,I}\right)$ in (3), a reduced fault tree model is generated by removing the events related to the relay failures and substituting the events associated with failures of all biosensors that are isolated due to the occurrence of $IE_{k,j}$ with constant "1" (meaning TRUE).

Step 5 Obtain the final WBAN unreliability by applying (3), (2) and (1) consecutively.

Fig. 2 gives the flowchart for the WBAN reliability evaluation based on the above procedure, where the operations that can be performed in parallel are illustrated. The operations in the shaded boxes for different values of k are independent and thus can also be performed in parallel given available computing resources.

5 Case Study

5.1 System Description

To demonstrate the reliability analysis method of Sect. 4, we consider an example of WBANs for patient monitoring in Fig. 3. The system contains four biosensors s_1, s_2,

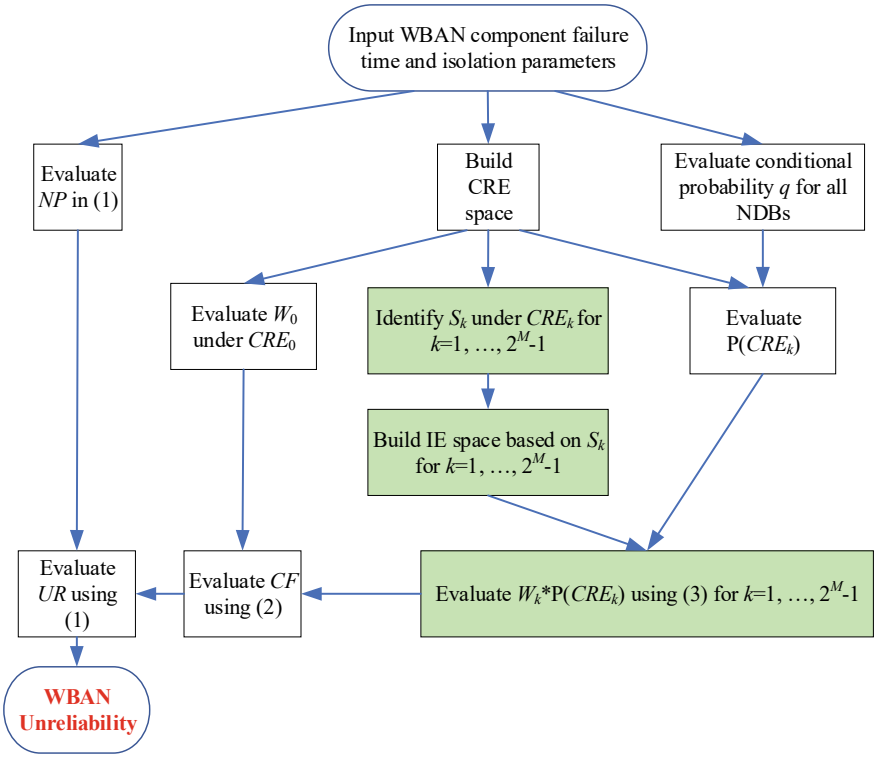


Fig. 2 Structure of a typical WBAN system

s_3, and s_4, which monitor the physiological information of the patient. Specifically, biosensors s_1 and s_2 are used to measure the patient's electromyogram (EMG) signals by detecting regular movement and muscle contraction; biosensors s_3 and s_4 are used to measure the patient's heart rate and blood pressure, respectively. The example WBAN system also contains a relay denoted by r, which, when functioning correctly, relays data sensed by s_1 and s_2 to the LPU.

The smart phone (i.e., LPU) receives and processes data from biosensors and the relay, facilitating health professionals to learn the functioning of nerves in the arm and the leg [42, 43] and other basic physiological information in response to a nerve's stimulation of the muscle.

The example WBAN system is reliable when the EMG, heart rate, and blood pressure data are all accessible by the LPU. In other words, the system fails when both s_1 and s_2 have failed, or s_3 fails, or s_4 fails. In this case study, the LPU is assumed to be fully reliable during the mission time and, thus, is not considered in the WBAN reliability analysis.

Figure 4 illustrates the fault tree model of the example WBAN system. The PFD gate in the fault tree models the PFD behavior that exists between relay r and biosensors s_1 and s_2. As discussed in Sect. 2, if relay r fails before PFs of biosensors s_1 and

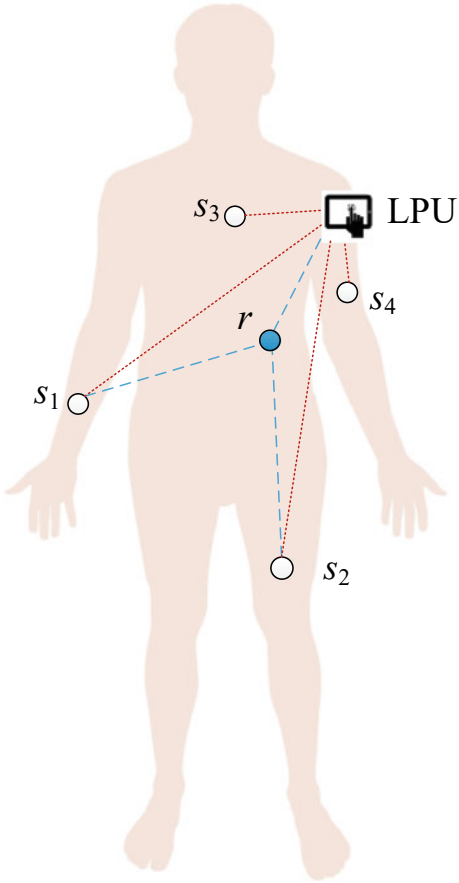


Fig. 3 An example of WBANs

s_2, s_1 and s_2 are isolated with certain probabilities, referred to as isolation factors and denoted by $IF_{r,s1}$ and $IF_{r,s2}$, respectively. The two isolation factors ($IF_{r,s1}$ and $IF_{r,s2}$) obey the Weibull distribution [19, 20] with the scale and shape parameters of (6.0e−3/month, 2) and (3.0e−3/month, 2), respectively.

Each biosensor in the example WBAN system is subject to LFs compromising its sensing function. Relay r is subject to LFs compromising its relaying/transmission function. Suppose that biosensors s_1, s_2, and s_4 are subject to jamming attacks (i.e., PFs). Since jamming attacks are based on transmitting illegal signals without compromising the sensing function, the LFs and PFs of biosensors s_1, s_2, and s_4 are s-independent events. Assuming time-to-LF and time-to-PF of WBAN components obey the Weibull distributions with scale parameter (λ/month) and shape parameter β [44]. Table 4 presents values of these parameters for the example WBAN components.

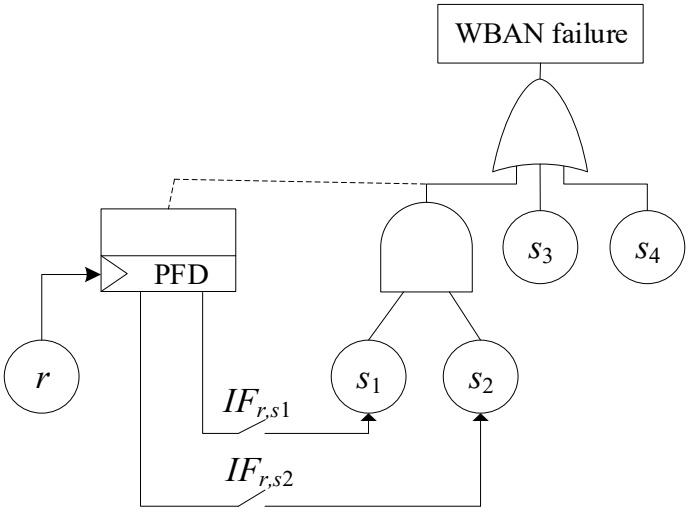


Fig. 4 Fault tree of the example WBAN system

Table 4 Weibull time-to-failure scale parameter (λ/month) and shape parameter β for WBAN components

WBAN component	LF	PF
r	(3.0e−3, 1)	–
s_1	(8.0e−4, 2)	(8.5e−2, 2)
s_2	(4.0e−4, 2)	(8.5e−2, 2)
s_3	(3.5e−4, 2)	–
s_4	(2.8e−4, 2)	(4.0e−2, 2)

5.2 Reliability Analysis

The reliability analysis procedure presented in Section 4 is illustrated using the example WBAN system. The single relay leads to two CREs: $CRE_0 = \overline{r}$ and $CRE_1 = r$. According to (2), we have

$$\begin{aligned} CF &= \sum_{k=0}^{2^1-1} \mathrm{P}(\text{WBAN fails}|CRE_k) \times \mathrm{P}(CRE_k) \\ &= W_0 \times \mathrm{P}(CRE_0) + W_1 \times \mathrm{P}(CRE_1) \end{aligned} \tag{4}$$

According to case 1 in Step 3, we evaluate $W_0 = \mathrm{P}(\text{WBAN fails}|CRE_0)$ by replacing event r with constant "0", leading to a reduced fault tree as shown in Fig. 5a. Using ordering of $s_3 \rightarrow s_4 \rightarrow s_1 \rightarrow s_2$, the reduced fault tree can then be converted to the BDD in Fig. 5b, which is then evaluated to obtain W_0. Refer to [41] for the BDD generation from a fault tree model and the BDD evaluation method.

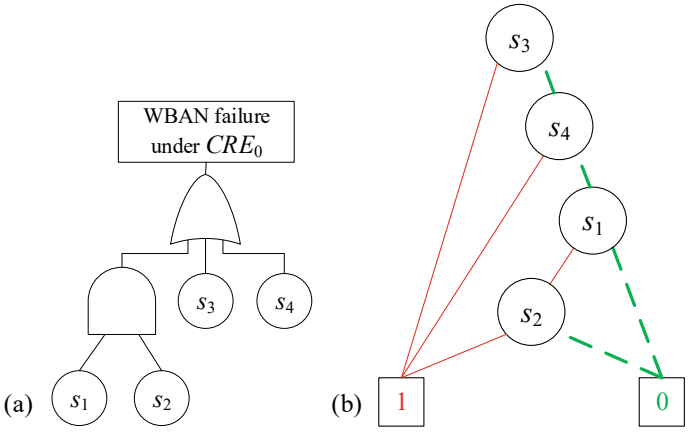


Fig. 5 Reliability models for evaluating W_0. **a** Reduced fault tree and **b** BDD

Under CRE_1 (i.e., the relay r fails), since there are two dependent biosensors s_1 and s_2, four IEs can be defined in Table 5. According to (3), we have

$$W_1 \times \mathrm{P}(CRE_1) = P(\text{WBAN fails}|CRE_1) \times P(CRE_1) = \sum_{j=0}^{3}\left[\begin{array}{c} P\left(IE_{1,j} \cap CRE_{1,P}\right)+ \\ P\left(\text{WBAN fails}|IE_{1,j} \cap CRE_{1,I}\right) \times P\left(IE_{1,j} \cap CRE_{1,I}\right) \end{array}\right] \tag{5}$$

Under $IE_{1,0}$, neither of the two dependent biosensors is isolated. Hence, P(WBAN fails|$IE_{1,0} \cap CRE_{1,I}$) in (5) can be evaluated using the fault tree and BDD in Fig. 5. Thus, we have P(WBAN fails|$IE_{1,0} \cap CRE_{1,I}$) = W_0.

Under $IE_{1,1}$, biosensor s_2 is isolated. According to Step 4 in Sect. 4, the reduced fault tree in Fig. 6a is generated by removing the PFD gate from the fault tree in Fig. 4 and replacing event s_2 with "1". The reduced fault tree is then converted to the BDD in Fig. 6b, which is evaluated to obtain P(WBAN fails|$IE_{1,1} \cap CRE_{1,I}$) in (5).

Under $IE_{1,2}$, biosensor s_1 is isolated. According to Step 4 in Sect. 4, the reduced fault tree in Fig. 7a is generated by removing the PFD gate from the fault tree in Fig. 4 and replacing event s_1 with "1". The reduced fault tree is then converted to the BDD in Fig. 7b, which is evaluated to obtain P(WBAN fails|$IE_{1,2} \cap CRE_{1,I}$) in (5).

Table 5 IE space under CRE_1

Notation	Definition	Explanation
$IE_{1,0}$	$\bar{s}_1 \cap \bar{s}_2$	Neither s_1 nor s_2 is isolated
$IE_{1,1}$	$\bar{s}_1 \cap s_2$	s_1 is not isolated, s_2 is isolated
$IE_{1,2}$	$s_1 \cap \bar{s}_2$	s_1 is isolated, s_2 is not isolated
$IE_{1,3}$	$s_1 \cap s_2$	s_1 and s_2 are both isolated

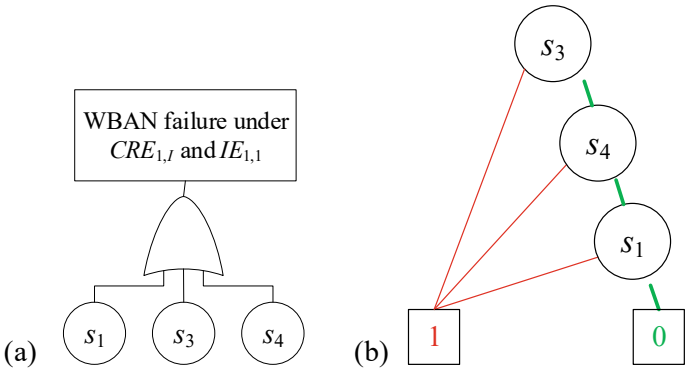


Fig. 6 Reliability models for evaluating P(WBAN fails|$IE_{1,1} \cap CRE_{1,I}$). **a** Reduced DFT, **b** BDD

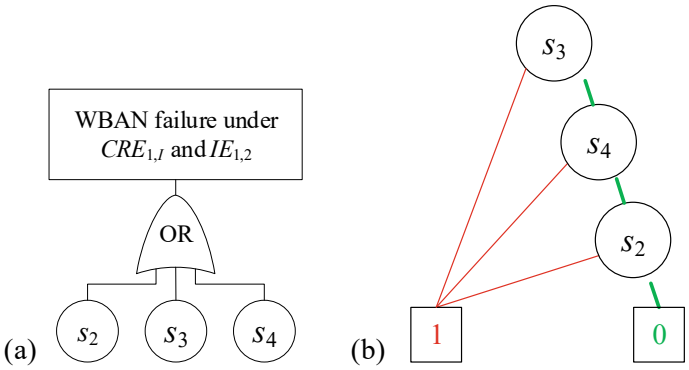


Fig. 7 Reliability models for evaluating P(WBAN fails|$IE_{1,2} \cap CRE_{1,I}$). **a** Reduced DFT, **b** BDD

Under $IE_{1,3}$, both s_1 and s_2 are isolated. In this case, after replacing events s_1 and s_2 in the fault tree of Fig. 4 with "1", the fault tree is reduced to "1", i.e., P(WBAN fails|$IE_{1,3} \cap CRE_{1,I}$) = 1.

Using input parameters given in Sect. 5.1, we obtain the unreliability of the example WBAN system at $t = 1$ day and $t = 1$ month as 0.000018 and 0.015922, respectively.

6 Summary

The fast-growing applications of WBANs in different critical domains pose high reliability requirements for WBAN designs and operations. This chapter conducts a critical review of reliability modeling, analysis, and design methods, providing

support for further research on WBAN reliability problems. A combinatorial reliability model is also presented, which offers an efficient solution to the reliability analysis of WBANs subject to the PFD and associated probabilistic isolation and competing behaviors. The model is demonstrated through a case study on a patient monitoring WBAN system.

References

1. Ramli SN, Ahmad R (2011) Surveying the wireless body area network in the realm of wireless communication. In: 2011 7th international conference on information assurance and security (IAS). IEEE, Melacca, Malaysia, pp 58–61
2. Jovanov E (2006) Wireless technology and system integration in body area networks for m-health applications. In: 2005 IEEE engineering in medicine and biology 27th annual conference. IEEE, Shanghai, China, pp 7158–7160
3. Fu Y, Ayyagari D, Colquitt N (2009) Pulmonary disease management system with distributed wearable sensors. In: 2009 annual international conference of the IEEE engineering in medicine and biology society. IEEE, USA, pp 773–776
4. Hall D, Llinas J (1998) From GI Joe to starship trooper: the evolution of information support for individual soldiers. In: ISCAS'98. Proceedings of the 1998 IEEE international symposium on circuits and systems (Cat. No. 98CH36187), vol 6, pp 541–544. IEEE, Monterey, CA, USA
5. Thakre L, Patil N, Kapse P, Potbhare P (2022) Implementation of soldier tracking and health monitoring system. In: 2022 10th international conference on emerging trends in engineering and technology-signal and information processing (ICETET-SIP-22). IEEE, Nagpur, India, pp 1–5
6. Ma X, Rupp E, Ryan A, Semper C, Thompson A, Zhang L (2014) Realtime physiological performance monitoring to prevent fatigue in special forces soldiers. In: 2014 40th annual northeast bioengineering conference (NEBEC). IEEE, Boston, MA, USA, pp 1–2
7. Hasan K, Biswas K, Ahmed K, Nafi NS, Islam MS (2019) A comprehensive review of wireless body area network. J Netw Comput Appl 143:178–198
8. Shokeen S, Parkash D (2019) A systematic review of wireless body area network. In: 2019 international conference on automation, computational and technology management (ICACTM). IEEE, London, UK, pp 58–62
9. Arefin MT, Ali MH, Haque AF (2017) Wireless body area network: an overview and various applications. J Comput Commun 5(7):53–64
10. Singh S, Prasad D (2022) Wireless body area network (WBAN): a review of schemes and protocols. Mater Today Proc 49:3488–3496
11. Al Barazanchi I, Hashim W, Alkahtani AA, Abbas HH, Abdulshaheed HR (2021) Overview of WBAN from literature survey to application implementation. In: 2021 8th international conference on electrical engineering, computer science and informatics (EECSI). IEEE, Semarang, Indonesia, pp 16–21
12. Gao GP, Yang C, Hu B, Zhang RF, Wang SF (2018) A wearable PIFA with an all-textile metasurface for 5 GHz WBAN applications. IEEE Antennas Wirel Propag Lett 18(2):288–292
13. El Atrash M, Abdalla MA, Elhennawy HM (2019) A wearable dual-band low profile high gain low SAR antenna AMC-backed for WBAN applications. IEEE Trans Antennas Propag 67(10):6378–6388
14. Dasgupta A, Mennemanteuil MM, Buret M, Cazier N, Colas-des-Francs G, Bouhelier A (2018) Optical wireless link between a nanoscale antenna and a transducing rectenna. Nat Commun 9(1):1–7
15. Grover K, Lim A, Yang Q (2014) Jamming and anti-jamming techniques in wireless networks: a survey. Int J Ad Hoc Ubiquitous Comput 17(4):197–215

16. Wang Y, Xing L, Wang H, Levitin G (2015) Combinatorial analysis of body sensor networks subject to probabilistic competing failures. Reliab Eng Syst Saf 142:388–398
17. Xing L, Levitin G, Wang C (2019) Dynamic system reliability: modeling and analysis of dynamic and dependent behaviors. Wiley, Hoboken, NJ
18. Xing L, Morrissette BA, Dugan JB (2014) Combinatorial reliability analysis of imperfect coverage systems subject to functional dependence. IEEE Trans Reliab 63(1):367–382
19. Wang Y, Xing L, Wang H, Coit DW (2017) System reliability modeling considering correlated probabilistic competing failures. IEEE Trans Reliab 67(2):416–431
20. Wang Y, Xing L, Levitin G, Huang N (2018) Probabilistic competing failure analysis in phased-mission systems. Reliab Eng Syst Saf 176:37–51
21. Zhao G, Xing L (2021) Reliability analysis of body sensor networks subject to random isolation time. Reliab Eng Syst Saf 207:107345
22. Xing L, Zhao G, Wang Y, Xiang Y (2020) Reliability modeling of correlated competitions and dependent components with random failure propagation time. Qual Reliab Eng Int 36(3):947–964
23. Wang S, Nah JW, Seak KJ, Park JT (2009) Analytical modeling of multi-type failures in wireless body area networks. In: 2009 2nd IEEE international conference on broadband network & multimedia technology. IEEE, Beijing, China, pp. 242–246
24. Divi K, Liu H (2013) Modeling of WBAN and cloud integration for secure and reliable healthcare. In Proceedings of the 8th international conference on body area networks. Boston, USA, pp 128–131
25. Strielkina A, Kharchenko V, Uzun D (2018) A Markov model of healthcare internet of things system considering failures of components. In: CEUR workshop proceedings. ICTERI, Kyiv, Ukraine, pp 530–543
26. Dehabadi MSZ, Jahed M (2017) Reliability modeling of anomaly detection algorithms for wireless body area networks. In 2017 Iranian conference on electrical engineering (ICEE). IEEE, Tehran, Iran, pp 70–75
27. Lomotey RK, Pry J, Sriramoju S (2017) Wearable IoT data stream traceability in a distributed health information system. Pervasive Mob Comput 40:692–707
28. Santos GL, Gomes D, Kelner J, Sadok D, Silva FA, Endo PT, Lynn T (2020) The internet of things for healthcare: optimising e-health system availability in the fog and cloud. Int J Comput Sci Eng 21(4):615–628
29. Troesch F, Steiner C, Zasowski T, Burger T, Wittneben A (2007) Hardware aware optimization of an ultra low power UWB communication system. In: 2007 IEEE international conference on ultra-wideband. IEEE, Singapore, pp 174–179
30. Masroor K, Jeoti V, Drieberg M (2018) Improving the energy efficiency of a wireless body area network using a redundant coordinator for healthcare applications. In: 2018 international conference on intelligent and advanced system (ICIAS). IEEE, Lumpur, Malaysia, pp 1–5
31. Awad M, Sallabi F, Shuaib K, Naeem F (2022) Artificial intelligence-based fault prediction framework for WBAN. J King Saud Univ-Comput Inform Sci 34(9):7126–7137
32. Arora N, Gupta SH, Kumar B (2022) Analyzing the best location for the central node placement for energy efficient and reliable WBAN. In: 2022 IEEE 11th international conference on communication systems and network technologies (CSNT). IEEE, Indore, India, pp 438–444
33. Umare A, Ghare P (2018) Optimization of routing algorithm for WBAN using genetic approach. In: 2018 9th international conference on computing, communication and networking technologies (ICCCNT). IEEE, Bengalura, India, pp 1–6
34. Mehmood G, Khan MZ, Abbas S, Faisal M, Rahman HU (2020) An energy-efficient and cooperative fault-tolerant communication approach for wireless body area network. IEEE Access 8:69134–69147
35. Salayma M, Al-Dubai A, Romdhani I, Nasser Y (2017) New dynamic, reliable and energy efficient scheduling for wireless body area networks (WBAN). In: 2017 IEEE international conference on communications (ICC). IEEE, Paris, France, pp 1–6
36. Xu YH, Yu G, Yong YT (2020) Deep reinforcement learning-based resource scheduling strategy for reliability-oriented wireless body area networks. IEEE Sens Lett 5(1):1–4

37. Peng Y, Wang X, Guo L, Wang Y, Deng Q (2017) An efficient network coding-based fault-tolerant mechanism in WBAN for smart healthcare monitoring systems. Appl Sci 7(8):817
38. Su X, Li C, Song Y, Yuan X (2017) WBAN on hardware: implementation and optimization based on IEEE 802.15. 6. In: 2017 international conference on computer, information and telecommunication systems (CITS). IEEE, Dalian, China, pp 147–151
39. Chen M, Li Z, Zhang G (2014) A cooperative software-hardware approach for wireless body area network implementation. In: The 4th annual IEEE international conference on cyber technology in automation, control and intelligent. IEEE, Hongkong, China, pp 214–218
40. Mathew P, Augustine L, Kushwaha D, Vivian D, Selvakumar D (2014) Hardware implementation of NB PHY baseband transceiver for IEEE 802.15. 6 WBAN. In: 2014 international conference on medical imaging, m-health and emerging communication systems (MedCom). IEEE, Greater Noida, India, pp 64–71
41. Xing L, Amari SV (2015) Binary decision diagrams and extensions for system reliability analysis. Wiley, Hoboken
42. Venuto DD, Annese VF, Tommaso MD, Vecchio E, Sangiovanni Vincentelli AL (2015) Combining EEG and EMG signals in a wireless system for preventing fall in neurodegenerative diseases. In: Ambient assisted living. Springer, Cham, pp 317–327
43. Al Rasyid MUH, Prasetyo D, Nadhori IU, Alasiry AH (2015) Mobile monitoring of muscular strain sensor based on wireless body area network. In: 2015 international electronics symposium (IES). IEEE, Surabaya, Indonesia, pp 284–287
44. Koc H, Shaik SS, Madupu PP (2019) Reliability modeling and analysis for cyber physical systems. In: 2019 IEEE 9th annual computing and communication workshop and conference (CCWC). IEEE, pp 0448–0451

Sensitivity Estimation of Markov Reward Models and Its Applications to Component Importance Analysis

Junjun Zheng, Hiroyuki Okamura, and Tadashi Dohi

Abstract Component importance analysis measures the effect on system reliability of components' reliabilities, enables the analyst to rank each component's contribution to the system failure, and identifies the system's weak components. Thus the system reliability can be improved by upgrading the weak components. Component importance analysis is commonly used in the design of a system from the reliability point of view. However, although dependencies exist among the failure behavior of systems in practice, and the dependent failures are known as a risk factor for degradation of system reliability, it is difficult to evaluate the component importance measures in the presence of failure dependencies analytically. In this chapter, we consider the Markov chain-based component-wise sensitivity analysis, which can evaluate the component importance measures without any system structure function. In particular, three types of component importance measures are derived from the viewpoints of both steady-state availability and reliability. Also, numerical examples illustrate the component importance analysis with the proposed approach.

Keywords Markov reward model · Failure dependency · Parametric sensitivity · Component importance analysis · Birnbaum importance · Criticality importance · Upgrading function

Abbreviations

CCF	Common-cause failure
SPN	Stochastic Petri net
CTMC	Continuous-time Markov chain
FT	Fault tree

J. Zheng (✉)
Osaka University, Osaka, Japan
e-mail: jzheng@ist.osaka-u.ac.jp

H. Okamura · T. Dohi
Hiroshima University, Hiroshima, Japan

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_5

RBD	Reliability block diagram
IIM	Integrated importance measure
ODE	Ordinary differential equation
MRM	Markov reward model
IB	Birnbaum component importance
AIB	Birnbaum availability component importance
RIB	Birnbuam reliability component importance
AICR	Availability criticality importance
RICR	Reliability criticality importance
AIU	Availability upgrading function
RIU	Reliability upgrading function

Notations

S	A denumerable state space
$\boldsymbol{X(t)}$	A stochastic process with the denumerable state space S
$\boldsymbol{Y(t)}$	A reward process
$\pi(t)$	State probability (row) vector at time t
π_{ss}	Steady-state probability vector
Q	Infinitesimal generator matrix
A	A matrix
I	Identity matrix
R	A column vector
$\boldsymbol{\theta}$	A model parameter
θ	Model parameter vector
$s(t, \theta)$	Sensitivity of state probability w.r.t. parameter θ at time t
$s_{ss}(\theta)$	Sensitivity of steady-state probability w.r.t. parameter θ
$S(\theta)$	Sensitivity of infinitesimal generator w.r.t. parameter θ
$\boldsymbol{I_S}$	Dependability/performance index of system
$\boldsymbol{I_k}$	Dependability/performance index of component k
r_S	Reward vector of system
r_k	Reward vector of component k
$\boldsymbol{R_S(t)}$	Reliability of system at time t
$\boldsymbol{R_k(t)}$	Reliability of component k at time t
$\boldsymbol{A_S}$	Steady-state availability of system
$\boldsymbol{A_k}$	Steady-state availability of component k
$\boldsymbol{\delta_j}$	Deviation of I_S with respect to θ_j that are not correlated to the deviations of $I_1, \ldots, I_K$
δ	The column vector whose elements are δ_j
u	The column vector whose elements are the sensitivities of system dependability/performance index w.r.t. the component performance indices

J	The matrix whose elements are the sensitivities of component dependability/performance indices w.r.t. each model parameter
z	The column vector whose elements are the sensitivities of system performance index w.r.t. each model parameter
$\|\delta\|_2$	2-Norm of vector $\vde$
T	The transpose operator
$\mathcal{U}_k$	The set of states where component k is up
$\mathcal{D}_k$	The set of states where component k is down
$\mathcal{U}_s$	The set of states where the system is up
$\mathcal{D}_s$	The set of states where the system is down
$[\cdot]_i$	The i th element of a vector
$[\cdot]_{i,j}$	The (i, j)-entry of a matrix
$\boldsymbol{x_k}$	A binary variable meaning the condition of component k
x	State vector of the system whose elements are x_k
$\phi(x)$	Structure function of the system
$\triangle\ \phi(k, x)$	The first derivative of structure function w.r.t. the state condition of component k
$P(x)$	A certain probability mass function for $\vx$
$\boldsymbol{\lambda_k}$	Failure rate of component k
$\boldsymbol{\mu_k}$	Repair rate of component k
$\boldsymbol{\lambda_k(t)}$	Time-dependent failure rate of component k at time t

1 Introduction

System reliability depends on the reliabilities of components; thus, it is important to know a thorough understanding of how each of the system components function and how these functions affect the system operation for improving and evaluating the reliability of a system. Component importance analysis is a quantification method of a system that measures the effect on system reliability of components' reliabilities. It enables the analyst to rank each component's contribution to the system failure and identify the weak components. In practice, system failure is often caused by the dependent failures among components, such as common-cause failures (CCFs), which are defined as any condition or event that affects several components inducing their simultaneous failures or malfunction [1]. The dependent failure is known as the factor of the degradation of system reliability and is usually caused when several processing units share a common software module. In other words, when a CCF occurs, all the components affected by the common-cause event will fail. Therefore, it is crucial to evaluate the effect of failure dependencies. In the past, some researchers have considered computer systems with failure dependencies. For example, Fricks and Trivedi [1] evaluated the effect of failure dependencies in reliability models developed by stochastic Petri nets (SPNs) and continuous-time Markov chains (CTMCs).

Assessing and improving the system reliability is always an important topic. Computer systems face the threat of reliability degradation due to aging problems. In order to achieve a highly reliable system, it is necessary to ensure that the critical components in the system are operational with high reliability. To detect the critical components in the system, sensitivity analysis [2] is effective. The sensitivity analysis is a method to estimate the magnitude of deviations of dependability/performance indices when system configuration changes. Generally, the parametric sensitivity is considered, which is the first derivative of the dependability/performance measure with respect to a model parameter. The parametric sensitivity can also be applied to optimizing system dependability/performance by combing mathematical programming and the evaluated effects on parameters. On the other hand, in reliability engineering, the component importance analysis is more preferred over the parametric sensitivity analysis. The component importance analysis, called the component-wise sensitivity analysis, is to estimate the first derivatives of reliability measures of the system with respect to the reliability measures of components. Thus, the component importance analysis can directly detect the critical components from the reliability point of view and has attracted widespread attention. For example, in [3], Fricks and Trivedi considered three types of component importance measures for Markov chains using structure function. The structure function represents the relationship between components' failures and system failure and can be obtained using symbolic analytical logic techniques, such as fault tree (FT) and reliability block diagram (RBD) analysis. Also, the authors [4] applied a novel component-wise sensitivity analysis to derive the availability upgrading functions under which components are statistically independent and described by general CTMCs. In [4], the presented method can derive the component importance measures only from a CTMC model without any system structure function.

In this chapter, we present a Markov chain-based component-wise sensitivity analysis approach to evaluate the component importance measures of computer systems modeled by MRMs in the presence of failure dependencies. Specifically, both transient and stationary analysis of Markov reward models (MRMs) are introduced first. Then based on the transient and stationary solutions, the parametric sensitivity of MRMs is presented. We finally propose the component-wise sensitivity analysis using the parametric sensitivity of MRMs and the component importance analysis using the component-wise sensitivity. In particular, three kinds of importance measures are considered.

The rest of this chapter is organized as follows. Section 2 reviews some related works. Section 3 introduces the fundamentals of CTMC and MRMs, and presents the parametric sensitivity of MRMs. In Sect. 4 we propose the MRM-based sensitivity estimation method and using which we present the component importance analysis in Sect. 5. Section 6 is devoted to validating the proposed approach with three systems. Finally, we conclude this paper with some remarks in Sect. 7.

2 Related Work

Birnbaum importance was first proposed by Birnbaum [5]. In [5], three types of importance measures were defined: structure importance, reliability importance, and lifetime importance measures. Later on, Lambert [6] introduced the criticality importance measure. Beeson and Andrews [7] extended the Birnbaum-time dependent lifetime importance to a noncoherent system by separately considering the contribution of the component failure to system failure and the contribution of component repair to system failure. Shen et al. [8] considered the Birnbaum importance for linear consecutive-k-out-of-n systems with sparse d using structure function. Concretely, they proposed two kinds of structure functions for Lin/Con/$k/n(d)$ and derived the ranking of Birnbaum importance of components for some different patterns for Lin/Con/$k/n(d)$. In [9], Si et al. presented the integrated importance measure (IIM). They focused on the expected number of component failures in unit time and evaluated how the transition of component states affects the system performance based on the probability distributions and transition rates of component states. Based on the presented IIM equations of component states, Dui et al. [10] investigated how the transition of component states affects system performance under the semi-Markov process.

On the other hand, coping with the failure dependencies in the system is challenging. Fricks et al. [1] evaluated the effect of failure dependencies in reliability models developed using SPNs and CTMCs. Kristiansen et al. [11] pointed out that although several approaches to construct component-based software reliability models have been proposed, most of these approaches tend to ignore the failure dependencies that usually exist among software components. From their works, it can be concluded that failure dependencies highly influence system reliability; therefore, the dependent failures should never be ignored. Several works have dealt with the component importance analysis for dependent failures system. For example, Pan and Nonaka [12] presented a quantitative method to evaluate the importance of each CCF event. Specifically, they divided the CCFs into two groups; one with a clear relationship between the causes and effects and the other without such relationship. In particular, the first group of CCFs evaluated the structure–function importance and probability importance of the common-root cause events modeled using FT. Also, they considered the Birnbaum importance for the second group of CCFs through a parametric model.

3 Preliminaries

3.1 Continuous-Time Markov Chain

Continuous-time Markov chain (CTMC) is a stochastic process with discrete state space on a continuous time domain. The CTMC is a convenient method to represent

the state transition of a system, such as normal and failure states, and thus is frequently used for the reliability evaluation of the system. In this chapter, we consider the parameter sensitivity of CTMC in both steady-state and transient analysis.

Le $\{X(t); t > 0\}$ be a time-homogeneous CTMC with the denumerable state space $S = \{1, 2, \ldots, n\}$ and $\boldsymbol{\pi}(t)$ is the state probability (row) vector whose i th element is the probability $P(X(t) = i)$. According to the fundamental CTMC analysis, we have

$$\frac{d}{dt}\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t)\mathbf{Q}, \tag{1}$$

where $\boldsymbol{Q}$ is an n-by-n square matrix called the infinitesimal generator. The (i, j)-entry of $\mathbf{Q}$ means the transition rate from state i to state j. Also, when 1 and 0 are column vectors where all the elements are 1 and 0, respectively, the diagonal entries of $\mathbf{Q}$ are given by the negative values such that $\mathbf{Q}1 = 0$.

In the transient analysis of CTMC, we focus on the probability vector $\boldsymbol{\pi}(t)$ at an arbitrary time under a given initial probability $\boldsymbol{\pi}(0) = \boldsymbol{\pi}_0$. In other words, the transient analysis solves the initial value problem of the ordinary differential equation (ODE) in Eq. (1) In general, by using the matrix exponential, the transient state probability vector can also be expressed by

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}_0 \exp(\mathbf{Q}t), \tag{2}$$

where the matrix exponential is defined by $\exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \mathbf{A}^2/2! + \cdots$ and $\mathbf{I}$ is an identity matrix. The uniformization is well-known as one of the most effective methods to solve the transient state probability vector [13].

On the other hand, the steady-state analysis of CTMC is to derive the steady-state probability vector; that is

$$\boldsymbol{\pi}_{ss}\mathbf{Q} = 0, \quad \boldsymbol{\pi}_{ss}1 = 1. \tag{3}$$

Intuitively, the steady-state probability vector $\boldsymbol{\pi}_{ss}$ corresponds to the (transient) state probability vector when $t \to \infty$. The steady-state probability vector can be obtained by solving the linear equation with GTH (Grassmann-Taksar-Heyman) algorithm, Gauss–Seidel, and SOR (successive over-relaxation).

3.2 Markov Reward Model

Markov reward models (MRMs) are a model-based approach for evaluating system dependability/performance. The MRM is generally defined by a CTMC and a reward function that maps the finite state space $\mathcal{S}$ to a real value. Let ρ be a reward function in an MRM. Then a rewarding process is given by $Y(t) = \rho(X(t))$ for the underlying CTMC process $X(t)$.

In fact, several kinds of reward functions have been discussed in the past literature [14]. Since this chapter deals with reliability and availability measures, we consider instantaneous reward functions in MRM. Let $\mathbf{r}$ be a column vector that maps each CTMC state $i \in \mathcal{S}$ to a corresponding real-valued reward r_i at time instance. Thus the expected instantaneous reward at time t is given by

$$\mathrm{E}[Y(t)] = \boldsymbol{\pi}(t)\mathbf{r}. \tag{4}$$

Also, the expected instantaneous reward in the steady state is also calculated by $\boldsymbol{\pi}_{ss}\mathbf{r}$.

3.3 Parametric Sensitivity of MRMs

The sensitivity analysis estimates the magnitude of deviations of dependability/performance indices when some parameters change. In particular, parametric sensitivity is the first or more derivatives of dependability/performance indices with respect to model parameters. The parametric sensitivity can also be applied to optimizing system dependability/performance by combing mathematical programming and the evaluated effects on parameters. In this section, we introduce the parametric sensitivity of general MRMs with instantaneous rewards. Similar to MRM analysis, the parametric sensitivity analysis is also divided into steady-state and transient cases.

Let θ be a model parameter of MRM. The parametric sensitivity analysis starts with computing the following sensitivity functions:

$$\mathbf{s}_{ss}(\theta) = \frac{\partial \boldsymbol{\pi}_{ss}}{\partial \theta}, \tag{5}$$

$$\mathbf{s}(t, \theta) = \frac{\partial \boldsymbol{\pi}(t)}{\partial \theta}. \tag{6}$$

If these sensitivity functions are obtained, the sensitivity of the dependability/performance index with instantaneous reward is given by

$$\frac{\partial}{\partial \theta}\boldsymbol{\pi}_{ss}\mathbf{r} = \mathbf{s}_{ss}(\theta)\mathbf{r} + \boldsymbol{\pi}_{ss}\frac{\partial}{\partial \theta}\mathbf{r}, \tag{7}$$

$$\frac{\partial}{\partial \theta}\mathrm{E}[Y(t)] = \mathbf{s}(t, \theta)\mathbf{r} + \boldsymbol{\pi}(t)\frac{\partial}{\partial \theta}\mathbf{r}. \tag{8}$$

Note that the above sensitivity functions of dependability/performance indices become simple when the reward vector is not sensitive to the parameter θ.

To obtain the sensitivity function in the case of a steady-state probability vector, we take the first derivative of Eq. (3);

$$\mathbf{s}_{ss}(\theta)\mathbf{Q} + \boldsymbol{\pi}_{ss}\mathbf{S}(\theta) = 0, \quad \mathbf{s}_{ss}(\theta)1 = 0, \tag{9}$$

where $\mathbf{S}(\theta) = \partial\mathbf{Q}/\partial\theta$. If the steady-state probability vector $\boldsymbol{\pi}_{ss}$ is already given, the sensitivity function $\mathbf{s}_{ss}(\theta)$ can be solved as the linear equation.

In the transient case, from Eq. (1), the sensitivity function holds the following ODE:

$$\frac{d}{dt}\mathbf{s}(t,\theta) = \mathbf{s}(t,\theta)\mathbf{Q} + \boldsymbol{\pi}(t)\mathbf{S}(\theta). \tag{10}$$

By integrating Eq. (1) into the above ODE, we have

$$\tfrac{d}{dt}\tilde{\pi}(t,\theta) = \tilde{\pi}(t,\theta)\tilde{\mathbf{Q}}(\theta), \tag{11}$$

where

$$\tilde{\pi}(t,\theta) = (\boldsymbol{\pi}(t), \mathbf{s}(t,\theta)), \quad \tilde{\mathbf{Q}}(\theta) = \begin{pmatrix} \mathbf{Q} & \mathbf{S}(\theta) \\ & \mathbf{Q} \end{pmatrix}. \tag{12}$$

Since the diagonal elements of $\tilde{\mathbf{Q}}(\theta)$ are the same as those of $\mathbf{Q}$, we can apply the uniformization to the following matrix exponential form:

$$\tilde{\pi}(t,\theta) = \tilde{\pi}(0,\theta)\exp\Big(\tilde{\mathbf{Q}}(\theta)t\Big). \tag{13}$$

4 Sensitivity Estimation Method

In this chapter, we consider the estimation method for the MRM-based dependability/performance model. As aforementioned, MRMs are a useful approach to evaluating the system dependability/performance based on CTMCs. Also, in Sect. 3.3, we introduce the parametric sensitivity of MRMs. However, we often encounter situations where parametric sensitivities are insufficient to investigate the system deviation. For example, the component importance analysis measures the deviation of the system reliability when components' reliabilities change. Except for some specific cases, such as s-independent components and an explicit structure function, it is not easy to obtain the sensitivity of component reliability on system reliability analytically.

Suppose that the system consists of K components. Let I_S and I_k be the dependability/performance indices of system and component k, respectively. This work assumes that the dependability/performance indices can be computed by an MRM with instantaneous rewards. Also, we define the reward vectors corresponding to I_S and I_k as $\mathbf{r}_S$ and $\mathbf{r}_k$, respectively. Without loss of generality, we assume

$$I_S = \pi \mathbf{r}_S, \tag{14}$$

$$I_k = \pi \mathbf{r}_k, \quad \text{for } k = 1, \ldots, K, \tag{15}$$

where $\boldsymbol{\pi}$ is a state probability vector of the underlying CTMC at an arbitrary time. In the case of transient measure, $\boldsymbol{\pi}$ should be $\boldsymbol{\pi}(t)$. On the other hand, $\boldsymbol{\pi}$ is the steady-state probability vector when the dependability/performance index is a steady-state measure.

In this chapter, we estimate the sensitivities of the system dependability/performance index with respect to component dependability/performance indices:

$$\frac{\partial I_S}{\partial I_k}, \quad \text{for } k = 1, \ldots, K. \tag{16}$$

As mentioned before, the above sensitivities cannot be obtained analytically, except for the case where I_S is explicitly given by a function of $I_1, \ldots, I_k$. Then we consider the estimation from the parametric sensitivities.

Consider the sensitivity of the system dependability/performance index with respect to model parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. According to the chain rule in partial differentiation, we have

$$\frac{\partial I_S}{\partial \theta_j} = \sum_{k=1}^{K} \frac{\partial I_S}{\partial I_k} \frac{\partial I_k}{\partial \theta_j} + \delta_j \tag{17}$$

for $k = 1, \ldots K$ and $j = 1, \ldots, m$

In the above, δ_j means the deviation of I_S with respect to θ_j, which are not correlated to the deviations of $I_1, \ldots, I_K$. The above equation can be rewritten as

$$\mathbf{z} = \mathbf{J}\mathbf{u} + \boldsymbol{\delta}, \tag{18}$$

where

$$\mathbf{z} = \begin{pmatrix} \frac{\partial I_S}{\partial \theta_1} \\ \vdots \\ \frac{\partial I_S}{\partial \theta_m} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \frac{\partial I_S}{\partial I_1} \\ \vdots \\ \frac{\partial I_S}{\partial I_K} \end{pmatrix}, \quad \boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix}, \tag{19}$$

$$\mathbf{J} = \begin{pmatrix} \frac{\partial I_1}{\partial \theta_1} & \cdots & \frac{\partial I_K}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial I_1}{\partial \theta_m} & \cdots & \frac{\partial I_K}{\partial \theta_m} \end{pmatrix}. \tag{20}$$

In the equation, the matrix $\mathbf{J}$ and the vector $\mathbf{z}$ can be computed from the parametric sensitivity approach described in Sect. 3.3. The problem now is reduced to solving the above linear equation with respect to $\mathbf{u}$. However, since the vector $\boldsymbol{\delta}$ is also undetermined, we should consider how to treat the vector $\boldsymbol{\delta}$.

If I_S is a function of only $I_1, \ldots, I_K$, then the vector $\boldsymbol{\delta}$ becomes 0 theoretically. On the other hand, even if I_S has non-zero $\boldsymbol{\delta}$, it is better to estimate the deviation of $\boldsymbol{\delta}$ as small as possible. For example, we suppose that the system dependability/performance index is given by

$$I_S = I_1 + I_2 - I_1 I_2. \tag{21}$$

In this case, it is natural to take the following sensitivities;

$$\frac{\partial I_S}{\partial I_1} = 1 - I_2, \quad \frac{\partial I_S}{\partial I_2} = 1 - I_1. \tag{22}$$

However, if $I_1 I_2$ is regarded as one variable, namely, $I_3 = I_1 I_2$, then the sensitivities become

$$\frac{\partial I_S}{\partial I_1} = 1, \quad \frac{\partial I_S}{\partial I_2} = 1, \quad \frac{\partial I_S}{\partial I_3} = -1. \tag{23}$$

Although Eqs. (22) and (23) are quite different, they provide the same results on the parametric sensitivity. That is, based on Eq. (22), we have

$$\begin{aligned} \frac{\partial I_S}{\partial \theta} &= \frac{\partial I_S}{\partial I_1}\frac{\partial I_1}{\partial \theta} + \frac{\partial I_S}{\partial I_2}\frac{\partial I_2}{\partial \theta} \\ &= (1 - I_2)\frac{\partial I_1}{\partial \theta} + (1 - I_1)\frac{\partial I_2}{\partial \theta}. \end{aligned} \tag{24}$$

When Eq. (23) is applied, the parametric sensitivity is expressed by

$$\begin{aligned} \frac{\partial I_S}{\partial \theta} &= \frac{\partial I_S}{\partial I_1}\frac{\partial I_1}{\partial \theta} + \frac{\partial I_S}{\partial I_2}\frac{\partial I_2}{\partial \theta} + \frac{\partial I_S}{\partial I_3}\frac{\partial I_3}{\partial \theta} \\ &= \frac{\partial I_1}{\partial \theta} + \frac{\partial I_2}{\partial \theta} - \frac{\partial I_3}{\partial \theta}. \end{aligned} \tag{25}$$

Therefore, from the mathematical point of view, both are correct. However, the former is a better representation of the relationship between the system dependability/performance index and component indices since $\partial I_S / \partial I_3$ can be expressed by $\partial I_S / \partial I_1$ and $\partial I_S / \partial I_2$. In other words, it is important to explain the deviation of I_S as much as possible by using only the deviations with I_1 and I_2. In Eq. (25), the last term corresponds to δ in Eq. (17). Therefore, it is better to take the estimates of $\mathbf{u}$ so that $\boldsymbol{\delta}$ becomes small.

Based on the above insight, we formulate the following mathematical programming:

$$\min_{\mathbf{u}} \|\boldsymbol{\delta}\|_2^2, \tag{26}$$

$$\text{s.t.} \quad \mathbf{z} = \mathbf{J}\mathbf{u} + \boldsymbol{\delta}, \tag{27}$$

where $\|\boldsymbol{\delta}\|_2$ is a 2-norm of vector $\boldsymbol{\delta}$. The problem is further reduced to the following least square problem:

$$\min_{\mathbf{u}} \|\mathbf{z} - \mathbf{J}\mathbf{u}\|_2^2. \tag{28}$$

There are several methods to solve the least square problem. The simplest approach is to solve the normal equation [15]:

$$\left(\mathbf{J}^T\mathbf{J}\right)\mathbf{u} = \mathbf{J}^T\mathbf{z}, \tag{29}$$

where T is the transpose operator. Then the estimates of sensitivities are given by

$$\mathbf{u} = \left(\mathbf{J}^T\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{z}. \tag{30}$$

5 Component Importance Analysis for MRMs

This section revisits the component importance analysis for MRMs with the sensitivity estimation method proposed in Sect. 4. Fricks and Trivedi [1] considered three component importance measures; Birnbaum importance measure, criticality measure, and upgrading function. In particular, they treated the MRM where components are statistically independent.

Suppose that the system consists of K components. In this chapter, the behavior of components is not completely separated, i.e., the system behavior with K components is only given by a CTMC with an infinitesimal generator $\mathbf{Q}$ with finite state space $\mathcal{S}$. On the other hand, the states of $\mathbf{Q}$ are classified into UP and DOWN states for both system and components. Let $\mathcal{U}_k$ and $\mathcal{D}_k$ be the sets of states where the component k is up and down, respectively. Also, $\mathcal{U}_S$ and $\mathcal{D}_S$ are the sets of states where the system is up and down. Note that $\mathcal{U}_k \cup \mathcal{D}_k = \emptyset, \mathcal{U}_S \cup \mathcal{D}_S = \emptyset, \mathcal{U}_k \cap \mathcal{D}_k = \mathcal{S}$ and $\mathcal{U}_S \cap \mathcal{D}_S = \mathcal{S}$. Then the reward vectors for component k and the system can be defined by

$$[r_k]_i = \begin{cases} 1, & i \in \mathcal{U}_k, \\ 0, & i \in \mathcal{D}_k, \end{cases} \tag{31}$$

and

$$[r_s]_i = \begin{cases} 1, & i \in \mathcal{U}_S, \\ 0, & i \in \mathcal{D}_S, \end{cases} \tag{32}$$

respectively, where $[\cdot]_i$ means the ith element of a vector. Using the reward vectors, the steady-state availabilities for component k and the system are

$$A_S = \boldsymbol{\pi}_{ss}\mathbf{r}_S, \quad A_k = \boldsymbol{\pi}_{ss}\mathbf{r}_k, \tag{33}$$

where $\boldsymbol{\pi}_{ss}$ is computed by solving Eq. (3).

On the other hand, if the underlying CTMC does not have transitions from $\mathcal{D}_k$ to $\mathcal{U}_k$ and from $\mathcal{D}_S$ to $\mathcal{U}_S$, the reliability functions of component k and the system are given by

$$R_S(t) = \boldsymbol{\pi}(t)\mathbf{r}_S, \quad R_k(t) = \boldsymbol{\pi}(t)\mathbf{r}_k, \tag{34}$$

where $\boldsymbol{\pi}(t)$ is obtained from Eq. (2).

5.1 Birnbaum Importance Measure

Let x_k be a binary variable representing the condition of component k;

$$x_k = \begin{cases} 1, & \text{if component } k \text{ is up,} \\ 0, & \text{if component } k \text{ is down.} \end{cases} \tag{35}$$

Besides, we define a vector $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ as a state vector of the system. The structure function represents the relationship between components' failures and system failure. In general, the structure function is defined by

$$\phi(\mathbf{x}) = \begin{cases} 1, & \text{if system is up,} \\ 0, & \text{if system is down.} \end{cases} \tag{36}$$

For example, if the system is a series system, namely, the system failure occurs when any component fails, the structure function is given by

$$\phi(\mathbf{x}) = x_1 x_2 \ldots x_K = \prod_{k=1}^{K} x_k. \tag{37}$$

If the system failure occurs only when all the components fail, the so-called parallel system, then the structure function is given by

$$\begin{aligned} \phi(\mathbf{x}) &= 1 - (1 - x_1)(1 - x_2)\ldots(1 - x_K) \\ &= 1 - \prod_{k=1}^{K}(1 - x_k). \end{aligned} \tag{38}$$

Birnbaum [5] defined the component importance from the reliability point of view. He first considered the first derivative of the structure function (namely, Birnbaum structure importance) with respect to the state condition of component k:

$$\Delta_\phi(k, \mathbf{x}) = \frac{\partial \phi(\mathbf{x})}{\partial x_k}. \tag{39}$$

Let $P(\mathbf{x})$ be a certain probability mass function for $\mathbf{x}$. Then the component importance, so-called Birnbaum component importance (IB), can be defined by

$$IB_k = \sum_{\mathbf{x}} \Delta_\phi(k, \mathbf{x}) P(\mathbf{x}). \tag{40}$$

Birnbaum availability and reliability component importance measures (AIB and RIB) become

$$AIB_k = \frac{\partial A_S}{\partial A_k}, \quad RIB_k(t) = \frac{\partial R_S(t)}{\partial R_k(t)}. \tag{41}$$

In particular, when components are statistically independent, AIB and RIB can be obtained from the structure function [1]:

$$AIB_k = \phi(1_k, \mathbf{A}) - \phi(0_k, \mathbf{A}), \tag{42}$$

$$RIB_k(t) = \phi(1_k, \mathbf{R}(t)) - \phi(0_k, \mathbf{R}(t)), \tag{43}$$

where, for K-dimensional vector $\mathbf{X}$,

$$(\alpha_k, \mathbf{X}) = (X_1, \ldots, X_{k-1}, \alpha, X_{k+1}, \ldots, X_K). \tag{44}$$

However, Eqs. (42) and (43) do not hold when components are dependent. Thus the estimation method presented in this work can be applied to computing $\partial A_S / \partial A_k$ and $\partial R_S(t) / \partial R_k(t)$ directly. Concretely, for the model parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, we compute

$$\mathbf{z}_A = \begin{pmatrix} \frac{\partial A_S}{\partial \theta_1} \\ \vdots \\ \frac{\partial A_S}{\partial \theta_m} \end{pmatrix}, \quad \mathbf{J}_A = \begin{pmatrix} \frac{\partial A_1}{\partial \theta_1} & \cdots & \frac{\partial A_K}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial A_1}{\partial \theta_m} & \cdots & \frac{\partial A_K}{\partial \theta_m} \end{pmatrix}, \tag{45}$$

$$\mathbf{z}_R(t) = \begin{pmatrix} \frac{\partial R_S(t)}{\partial \theta_1} \\ \vdots \\ \frac{\partial R_S(t)}{\partial \theta_m} \end{pmatrix}, \quad \mathbf{J}_R(t) = \begin{pmatrix} \frac{\partial R_1(t)}{\partial \theta_1} & \cdots & \frac{\partial R_K(t)}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial R_1(t)}{\partial \theta_m} & \cdots & \frac{\partial R_K(t)}{\partial \theta_m} \end{pmatrix}. \tag{46}$$

Then the estimates of $AIB = (AIB_1, \ldots, AIB_K)^{\mathrm{T}}$ and $RIB(t) = (RIB_1(t), \ldots, RIB_K(t))^{\mathrm{T}}$ are given by

$$AIB = \left(\mathbf{J}_A^T \mathbf{J}_A\right)^{-1} \mathbf{J}_A^T \mathbf{z}_A, \tag{47}$$

$$RIB(t) = \left(\mathbf{J}_R(t)^T \mathbf{J}_R(t)\right)^{-1} \mathbf{J}_R(t)^T \mathbf{z}_R(t). \tag{48}$$

5.2 Criticality Importance Measure

The criticality measure [16] means the probability that, when the system fails, the failure of component k becomes a cause of the system failure. According to Fricks and Trivedi [1], the criticality importance measures of availability and reliability (AICR and RICR) can be derived by

$$AICR_k = \frac{A_k}{A_S} \frac{\partial A_S}{\partial A_k}, \tag{49}$$

$$RICR_k(t) = \frac{R_k(t)}{R_S(t)} \frac{\partial R_S(t)}{\partial R_k(t)}. \tag{50}$$

Essentially, these measures can be computed from AIB_k and RIB_k, i.e.,

$$AICR_k = \frac{A_k}{A_S} AIB_k, \quad RICR_k(t) = \frac{R_k(t)}{R_S(t)} RIB_k(t). \tag{51}$$

5.3 Upgrading Function

The upgrading function is the parametric sensitivity function with respect to a failure rate [1]. According to the definition, we have the availability and reliability upgrading functions (AIU and RIU) for component k:

$$AIU_{k,\lambda} = \frac{\lambda_k}{A_S} \frac{\partial A_S}{\partial \lambda_k}, \tag{52}$$

$$AIU_{k,\mu} = \frac{\mu_k}{A_S} \frac{\partial A_S}{\partial \mu_k}, \tag{53}$$

$$RIU_{k,\lambda}(t) = \frac{\lambda_k}{R_S(t)} \frac{\partial R_S(t)}{\partial \lambda_k}, \tag{54}$$

where λ_k and μ_k are failure and repair rates of component k. Note that AIU can also be defined for the repair rate. The AIU is essentially the same as the availability importance measures discussed by Cassady et al. [17].

In [17] and [18], it was assumed that components are independent and each has only two states; that is, up and down on the underlying CTMC. On the other hand, the authors introduced the method to derive the availability upgrading functions under which components are described by general CTMCs. The idea behind the approach is to apply the aggregation technique [19]. However, components were assumed to be independent even in [4]. In this chapter, by applying sensitivity estimation and aggregation, we derive AIUs and RIU for MRMs.

First, we consider AIUs. Aggregation is a technique to reduce MRM-based availability models to the 2-state model, which has the same availability as the original model. When we focus on the state of one component, the states can be classified into $\mathcal{U}_k$ and $\mathcal{D}_k$. The aggregation technique converts the original model to the 2-state model with transitions from up to down states and up to down state. By applying this technique, we obtain failure and repair rates in the steady state that ensure the steady-state probabilities of the up (down) states keep the same as those in the original model. The obtained failure and repairs in the steady state are called *equivalent failure and repair rates* [19].

From the argument of CTMC, the equivalent failure and repair rates of component k can be computed as follows.

$$\lambda_k = \frac{\sum_{(i,j)\in\mathcal{U}_k\times\mathcal{D}_k}[\pi_{ss}]_i[\mathrm{Q}]_{i,j}}{\sum_{i\in\mathcal{U}_k}[\pi_{ss}]_i}, \tag{55}$$

$$\mu_k = \frac{\sum_{(j,i)\in\mathcal{D}_k\times\mathcal{U}_k}[\pi_{ss}]_j[\mathrm{Q}]_{j,i}}{\sum_{j\in\mathcal{D}_k}[\pi_{ss}]_j}, \tag{56}$$

where $[\cdot]_{i,j}$ is an (i, j)-entry of a matrix. By taking account of $A_k = \mu_k/(\lambda_k + \mu_k)$, $AIU_{k,\lambda}$ can be rewritten by

$$\begin{aligned} AIU_{k,\lambda} &= \frac{\lambda_k}{A_S}\frac{\partial A_S}{\partial \lambda_k} \\ &= \frac{\lambda_k}{A_S}\sum_{l=1}^{K}\frac{\partial A_S}{\partial A_l}\frac{\partial A_l}{\partial \lambda_k} \\ &= \frac{\lambda_k}{A_S}\frac{\partial A_S}{\partial A_k}\frac{\partial A_k}{\partial \lambda_k} \\ &= \frac{\lambda_k}{A_S}\frac{\partial A_S}{\partial A_k}\left(-\frac{\mu_k}{(\lambda_k+\mu_k)^2}\right) \\ &= -\frac{A_k}{A_S}\frac{\partial A_S}{\partial A_k}\frac{\lambda_k}{\lambda_k+\mu_k} \\ &= -\frac{A_k}{A_S}\frac{\partial A_S}{\partial A_k}\frac{\lambda_k}{\lambda_k+\mu_k} \end{aligned} \tag{57}$$

Similarly, $AIU_{k,\mu}$ becomes

$$\begin{aligned} AIU_{k,\mu} &= \frac{\mu_k}{A_S}\frac{\partial A_S}{\partial \mu_k} \\ &= \frac{\mu_k}{A_S}\frac{\partial A_S}{\partial A_k}\frac{\partial A_k}{\partial \mu_k} \\ &= \frac{\mu_k}{A_S}\frac{\partial A_S}{\partial A_k}\frac{\lambda_k}{(\lambda_k+\mu_k)^2} \\ &= \frac{A_k}{A_S}\frac{\partial A_S}{\partial A_k}\frac{\lambda_k}{\lambda_k+\mu_k} \\ &= (1-A_k)AICR_k. \end{aligned} \tag{58}$$

Next, we consider RIU. In this case, the equivalent failure rate cannot be computed. Instead we use the time-dependent failure rate, i.e., $\lambda_k(t) = -(dR_k(t)/dt)/(1-R_k(t))$. Generally, the relationship between the reliability function and the failure rate is given by

$$R_k(t) = e^{-\int_0^t \lambda_k(s)ds}. \tag{59}$$

Based on the above, RIU can be obtained by

$$\begin{aligned} RIU_{k,\lambda} &= \frac{\lambda_k(t)}{R_S(t)}\frac{\partial R_S(t)}{\partial \lambda_k(t)} \\ &= \frac{\lambda_k(t)}{R_S(t)}\sum_{l=1}^{K}\frac{\partial R_S(t)}{\partial R_l(t)}\frac{\partial R_l(t)}{\partial \lambda_k(t)} \\ &= \frac{\lambda_k(t)}{R_S(t)}\frac{\partial R_S(t)}{\partial R_k(t)}\frac{\partial R_k(t)}{\partial \lambda_k(t)} \\ &= \frac{\lambda_k(t)}{R_S(t)}\frac{\partial R_S(t)}{\partial R_k(t)}(-tR_k(t)) \\ &= -t\lambda_k(t)\frac{R_k(t)}{R_S(t)}\frac{\partial R_S(t)}{\partial R_k(t)} \\ &= -t\lambda_k(t)RICR_k(t). \end{aligned} \tag{60}$$

In the MRM, the failure rate of component k is

$$\lambda_k(t) = -\frac{\boldsymbol{\pi}(t)\mathbf{Q}\mathbf{r}_k}{\boldsymbol{\pi}(t)\mathbf{r}_k}. \tag{61}$$

6 Numerical Examples

In this section, we illustrate the component importance analysis using classic (structure function-based) and proposed (CTMC-based component-wise sensitivity analysis) methods for the systems with no dependent failures and the proposed method for solving the cases with failure dependencies. The following systems are considered.

- Two-unit hot standby system.
- A TMR (triple modular redundancy) system.
- A 1-out-of-3 system.

Particularly, for the hot standby system and TMR system, we consider both independent and dependent components cases as below:

- Assumption 1: Components are statistically independent.
- Assumption 2: Components are failure-dependent.

6.1 Two-Unit Hot Standby System

Consider a standby system consisting of two units a and b, each with failure rates $\lambda_a = 0.001$ [h^{-1}] and $\lambda_b = 0.002$ [h^{-1}], respectively, and repair rates μ_a and μ_b that are all 1 [h^{-1}]. Let a be the primary unit and b be the spare one. Suppose that there is a perfect fault-detection unit that detects failures in the primary unit and replaces it with the spare one, and each one can be repaired first when all units have failed. If both primary and spare units are powered up, thus spare can be switched into use immediately after the primary unit has failed. In such a case, the system is called hot standby redundancy. This subsection considers the two-unit hot standby system.

6.1.1 Assumption 1: Components Are Statistically Independent

Figure 1 shows the RBD of the hot standby system. The dynamics of the system are modeled by a CTMC (see Fig. 2). The CTMC model defines a system state as a combination of operating and failed components. Table 1 enumerates different possible states, where "O" indicates an operational component, and "F" indicates a failed one.

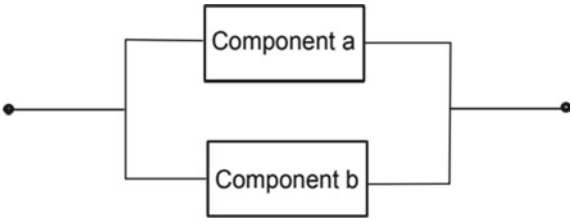


Fig. 1 RBD of two-unit hot standby system in a simple parallel configuration

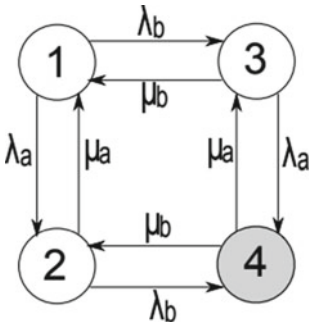


Fig. 2 State transition diagram for two-unit hot standby system

Table 1 The states of the standby systems

Component		State number
Primary (a)	Spare (b)	
O	O	1
F	O	2
O	F	3
F	F	4

Let A_i and $R_i(t)$ be the steady-state availability and reliability of component i, $i \in \{a, b\}$, respectively. Thus according to the parallel configuration in Fig. 1, we have the structure function for the system availability

$$\begin{aligned} A_{hs} &= 1 - (1 - A_a)(1 - A_b) \\ &= A_a + A_b - A_a A_b, \end{aligned} \tag{62}$$

where component availability $A_i = \mu_i/(\mu_i + \lambda_i)$. In the same manner, the structure function for the system reliability is given by

$$\begin{aligned} R_{hs}(t) &= 1 - (1 - R_a(t))(1 - R_b(t)) \\ &= R_a(t) + R_b(t) - R_a(t)R_b(t). \end{aligned} \tag{63}$$

In the above, the reliability of component i is computed by $R_i(t) = e^{-\lambda_i t}$. Thus the Birnbaum availability importance of components i becomes

$$AIB_i = \frac{\partial A_{hs}}{\partial A_i} = \begin{cases} 1 - A_b, & i = a \\ 1 - A_a, & i = b. \end{cases} \tag{64}$$

On the other hand, the Birnbaum reliability component importance measure $RIB_i(t)$ can be obtained similarly to Eq. (64). The availability and reliability of the system are computed using Markov analysis. For example, for availability evaluation, we should be able to repair the system from a failed state. The corresponding

infinitesimal generator is given by

$$\mathbf{Q}_{hs} = \begin{pmatrix} -(\lambda_a + \lambda_b) & \lambda_a & \lambda_b & 0 \\ \mu_a & -(\mu_a + \lambda_b) & 0 & \lambda_b \\ \mu_b & 0 & -(\mu_b + \lambda_a) & \lambda_a \\ 0 & \mu_b & \mu_a & -(\mu_b + \mu_a) \end{pmatrix}. \tag{65}$$

Then the steady-state probability vector of the hot standby system $\boldsymbol{\pi}_{hs}$ is obtained by solving the following linear equations:

$$\boldsymbol{\pi}_{hs}\mathbf{Q}_{hs} = 0, \quad \boldsymbol{\pi}_{hs}1 = 1. \tag{66}$$

Using structure function and component-wise sensitivity analysis based on the CTMC, respectively, we obtain the same results shown in Tables 2 through 4. In Table 2, we find that both the availability and reliability of component b are lower than those of component a since the failure rates $\lambda_b > \lambda_a$.

Table 3 demonstrates the availability importance measures, whereas Table 4 gives the reliability importance measures. Concretely, in Table 3, the Birnbaum importance of component a is higher than that of component b. This is because according to Eq. (64), the Birnbaum importance of a is determined by the availability of b and vice versa, and from Table 2 a has higher availability than b. Besides, the measure AICR is computed based on AIB, and AIU is given by AICR so that both AICR and AIU have the same trends as AIB. Similar results appear in the case of reliability importance.

Table 2 Availabilities and reliabilities ($t = 50$ h) of components and hot standby system

Component	Availability	Reliability
a	0.999001	0.951229
b	0.998004	0.904837
System	0.999998	0.995359

Table 3 Availability importance measures of components in the hot standby system

Component	AIB	$AICR$	AIU_λ	AIU_μ
a	1.996e−3	1.994e−3	−1.992e−6	1.992e−6
b	9.990e−4	9.970e−4	−1.990e−6	1.990e−6

Table 4 Reliability importance measures of components in the hot standby system ($t = 50$ h)

Component	$RIB(50)$	$RICR(50)$	$RIU_\lambda(50)$
a	9.516e−2	9.094e−2	−4.547e−3
b	4.877e−2	4.434e−2	−4.434e−3

6.1.2 Assumption 2: Components Are Failure-Dependent

In practice, computer systems often fail due to dependent failures. Suppose that either of two units can cause the concurrent failure of a and b in the hot standby system. Figure 3 shows the state transition diagram for the hot standby system with failure dependency. Note that the transition corresponding to the dependent failure is drawn as a dashed line. The figure shows that the two units' concurrent failures are modeled using the β factor method [20]. In the model, we assume that $\beta = 0.02$.

Table 5 demonstrates the time-dependent failure rate of components in the hot standby system at time $t = 50$ h. We see that under the dependent failure case (assumption 2), the failure rates of all components increase over time, which causes a decrease in both availability and reliability as seen in Table 6, compared with the results in Table 2.

The availability and reliability importance measures are given in Tables 7 and 8. The values of most importance measures become smaller in the system with failure dependency compared with the case with no dependent failures.

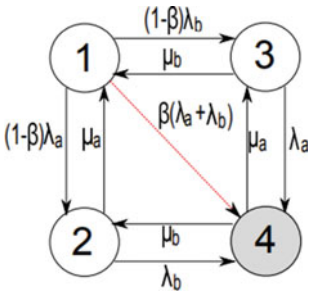


Fig. 3 State transition diagram for two-unit hot standby system with failure dependency

Table 5 Time-dependent failure rates of components in hot standby system ($t = 50$ h)

Component	Initial failure rate	Time-dependent failure rate	
		Under assumption 1	Under assumption 2
a	0.00100	0.00100	0.00104
b	0.00200	0.00200	0.00202

Table 6 Availabilities and reliabilities ($t = 50$ h) of components and hot standby system with failure dependency

Component	Availability	Reliability
a	0.998961	0.949419
b	0.997984	0.903955
System	0.999968	0.992666

Table 7 Availability importance measures of components in the hot standby system with failure dependency

Component	AIB	$AICR$	AIU_λ	AIU_μ
a	1.964e−3	1.962e−3	−2.038e−6	2.038e−6
b	9.756e−4	9.737e−4	−1.963e−6	1.963e−6

Table 8 Reliability importance measures of components in the hot standby system with failure dependency ($t = 50$ h)

Component	$RIB(50)$	$RICR(50)$	$RIU_\lambda(50)$
a	9.300e−2	8.895e−2	−4.609e−3
b	4.695e−2	4.276e−2	−4.316e−3

6.2 *Triple Modular Redundancy System*

Next, we consider the component importance analysis for a triple modular redundancy (TMR) system. In this system, three non-identical redundant processors carry out the same task simultaneously, and a voter compares the outputs from all the processors and sides with the majority. The system fails only when two or more processors fail, or the voter fails. In other words, the system can tolerate the failure of a single processor. A block diagram of the TMR system with a voter is shown in Fig. 4, and the corresponding RBD representation is depicted in Fig. 5. From these figures, the system is divided into processor and voter subsystems that are described by CTMCs.

6.2.1 Assumption 1: Components Are Statistically Independent

Figures 6 and 7 separately illustrate the CTMC for processor and voter subsystems. In Fig. 6, white and gray nodes represent the operational and failure states of the

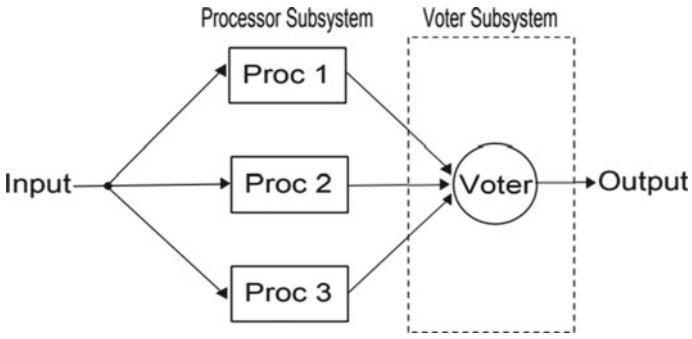


Fig. 4 Block diagram for TMR system with a voter

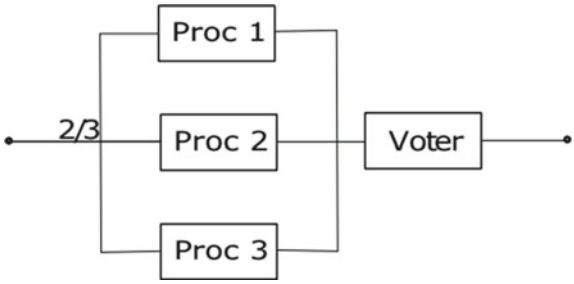


Fig. 5 RBD representation for TMR system

processor subsystem, respectively. More specifically, each state is indicated by three characters. The first character means the state of processor 1. When processor 1 is operational, the character is given by '1', and if failed, it is '0'. The second and third characters represent the states of processor 2 and processor 3, respectively, in the same manner as the first character. In the CTMC model for the voter, state UP means that the voter is operational. If the voter fails, it is denoted by state DN. Also, the model parameters are described in Table 9.

Let Q_p and Q_v be the infinitesimal generators of CTMCs for the processor subsystem and component voter, respectively. Then we have the composite CTMC generator for the TMR system by using the tensor sum of matrices [21] as

$$Q_S = Q_p \oplus Q_v. \tag{67}$$

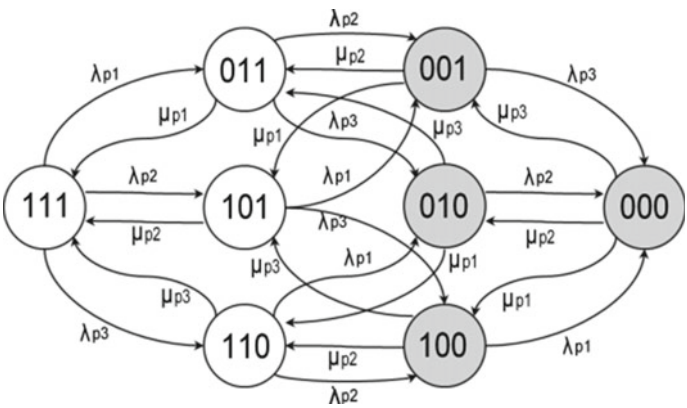


Fig. 6 CTMC of processor subsystem

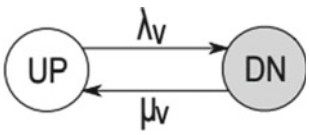


Fig. 7 CTMC of the component voter

Table 9 Model parameters for TMR system

Parameter	Description	Value (h)
$1/\lambda_{p1}$	Mean time to processor 1 failure	1250
$1/\lambda_{p2}$	Mean time to processor 2 failure	1000
$1/\lambda_{p3}$	Mean time to processor 3 failure	500
$1/\lambda_{v}$	Mean time to voter failure	10, 000
$1/\mu_{p1}$	Mean time to repair processor 1	1
$1/\mu_{p2}$	Mean time to repair processor 2	1
$1/\mu_{p3}$	Mean time to repair processor 3	1
$1/\mu_{v}$	Mean time to repair voter	4

Let A_i, and $R_i(t)$ be the steady-state availability and reliability of processor i, $i \in \{1, 2, 3\}$, and A_v and $R_v(t)$ be the steady-state availability and reliability of voter, respectively. Thus the availabilities and reliabilities of each component and system can be computed by using from Eqs. (31) to (34). On the other hand, according to RBD analysis (see Fig. 5), the system availability also can be obtained by

$$\begin{aligned} A_{tmr} &= (A_1A_2A_3 + (1 - A_1)A_2A_3 + A_1(1 - A_2)A_3 + A_1A_2(1 - A_3))A_v \\ &= (A_1A_2 + A_1A_3 + A_2A_3 - 2A_1A_2A_3)A_v. \end{aligned} \tag{68}$$

Similarly, the reliability of the TMR system is given by

$$\begin{aligned} R_{tmr}(t) &= (R_1(t)R_2(t)R_3(t) + (1 - R_1(t))R_2(t)R_3(t) + R_1(t)(1 - R_2(t))R_3(t) \\ &+ R_1(t)R_2(t)(1 - R_3(t))R_v(t) = (R_1(t)R_2(t) + R_1(t)R_3(t) \\ &+ R_2(t)R_3(t) - 2R_1(t)R_2(t)\ R_3(t))R_v(t). \end{aligned} \tag{69}$$

Table 10 shows the availabilities and reliabilities ($t = 50$ h) of components and system. In the table, the system availability and reliability are computed by using Eqs. (68) and (69). Processor 1 has the highest availability/reliability among the three processors because it has the longest MTTF (mean time to failure). In addition, the voter is more reliable than the processors.

Based on Eqs. (68) and (69), the Birnbaum availability and reliability importance of each component can be easily derived in the same manner as in the hot standby

Table 10 Availabilities and reliabilities ($t = 50$ h) of components and TMR system

Component	Availability	Reliability
Proc 1	0.999201	0.960789
Proc 2	0.999001	0.951229
Proc 3	0.998004	0.904837
Voter	0.999600	0.995012
System	0.999596	0.985141

Table 11 Availability importance measures of components in the TMR system

Component	AIB	$AICR$	AIU_{λ}	AIU_{μ}
Proc 1	2.990e−3	2.989e−3	−2.389e−6	2.389e−6
Proc 2	2.791e−3	2.789e−3	−2.787e−6	2.787e−6
Proc 3	1.796e−3	1.793e−3	−3.579e−6	3.579e−6
Voter	9.999e−1	1.000e−0	−3.998e−4	3.998e−4

Table 12 Reliability importance measures of components in the TMR system ($t = 50$ h)

Component	$RIB(50)$	$RICR(50)$	$RIU_{\lambda}(50)$
Proc 1	1.340e−1	1.307e−1	−5.227e−3
Proc 2	1.263e−1	1.219e−1	−6.097e−3
Proc 3	8.374e−2	7.691e−2	−7.691e−3
Voter	9.901e−1	1.000e−0	−5.000e−3

system and are given in Tables 11 and 12. The same results in these tables can also be obtained by using CTMC-based component-wise sensitivity analysis. From these tables, we see that the voter is the most important component contributing to system availability and reliability. On the other hand, processor 3 is the least important in terms of Birnbaum and criticality importance measures.

6.2.2 Assumption 2: Components Are Failure-Dependent

Consider the CCFs in the processor subsystem. The state transition diagrams of the processor subsystem with failure dependencies are illustrated in Fig. 8. The model parameters are also shown in Table 9. Here the β factor is also given as 0.02.

Table 13 demonstrates the time-dependent failure rates of components in the TMR system at time $t = 50$ h. Due to the impact of dependent failures, all processors' failure rates increase over time.

Table 14 shows the availabilities and reliabilities at time $t = 50$ h of components and TMR system with failure dependencies. Although the component voter is still reliable, as in the case without dependent failures, the system availability/reliability is decreased due to the increased failure risk brought by the failure dependencies.

The availability and reliability importance measures of components in the TMR system with failure dependencies are represented in Tables 15 and 16, respectively. From these tables, it is observed that in either availability or reliability aspect, processors become more important, compared with the case that components are statistically independent. On the other hand, the importance of the voter is still the highest and changes not much under different assumptions.

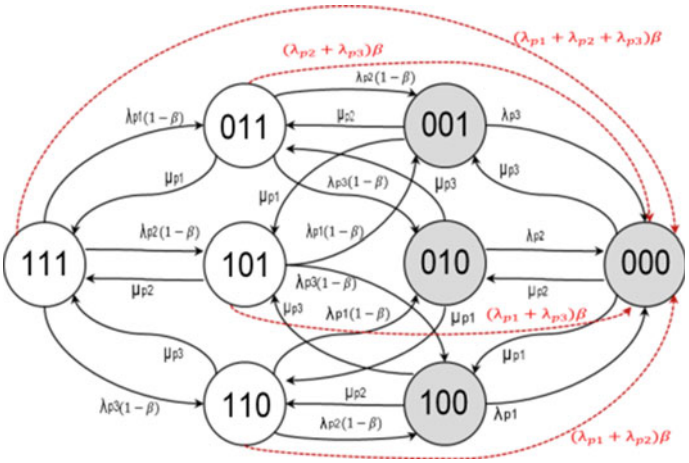


Fig. 8 CTMC for processor subsystem with failure dependencies

Table 13 Time-dependent failure rates of components in TMR system ($t = 50$ h)

Component	Initial failure rate	Time-dependent failure rate	
		Under assumption 1	Under assumption 2
Proc 1	0.00080	0.00080	0.00086
Proc 2	0.00100	0.00100	0.00105
Proc 3	0.00200	0.00200	0.00203
Voter	0.00010	0.00010	0.00010

Table 14 Availabilities and reliabilities ($t = 50$ h) of components and TMR system with failure dependencies

Component	Availability	Reliability
Proc 1	0.999141	0.958025
Proc 2	0.998945	0.948674
Proc 3	0.997968	0.903246
Voter	0.999600	0.995012
System	0.999533	0.981895

Table 15 Availability importance measures of components in TMR system with failure dependencies

Component	AIB	$AICR$	AIU_λ	AIU_μ
Proc 1	1.887e−2	1.887e−2	−1.621e−5	1.621e−5
Proc 2	1.868e−2	1.867e−2	−1.969e−5	1.969e−5
Proc 3	1.770e−2	1.767e−2	−3.590e−5	3.590e−5
Voter	9.999e−1	1.000e−0	−3.998e−4	3.998e−4

Table 16 Reliability importance measures of components in TMR system with failure dependencies ($t = 50$ h)

Component	$RIB(50)$	$RICR(50)$	$RIU_{\lambda}(50)$
Proc 1	1.445e−1	1.410e−1	−6.030e−3
Proc 2	1.369e−1	1.323e−1	−6.957e−3
Proc 3	9.504e−2	8.743e−2	−8.894e−3
Voter	9.868e−1	1.000e−0	−5.000e−3

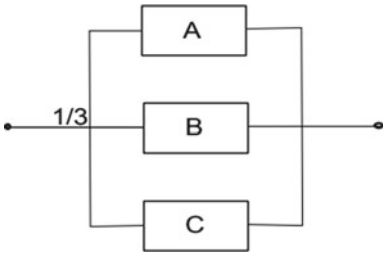


Fig. 9 RBD of 1-out-of-3 system

6.3 1-Out-of-3 System

In the case of a system with failure dependencies, we can use the proposed method to compute the component importance. However, suppose we just need to obtain the importance ranking of components and do not care about the importance values. In that case, we can also obtain the importance ranking from the structure function. We next consider a 1-out-of-3 system whose RBD is shown in Fig. 9 as an illustrative example. The system consists of three identical components A, B, and C, with failure rates $\lambda_A = \lambda_B = \lambda_C = 0.001$ [h^{-1}], and operates if at least one component operates.

6.3.1 Reliability Function

In the β-factor model, let $q_i(t)$, $i \in \{A, B, C\}$ be the independent failure probability of component i at time t. Also, let event E be the concurrent failure of components A, B, and C due to the failure of A. The occurrence probability of E is given by

$$P(E) = P(E|A)P(A) = \beta P(A) = \beta q_A(t) = q_E(t), \tag{70}$$

where β is given as 0.002. Then we have the system failure probability:

$$\begin{aligned} P(SF)(t) &= P(A)\left(P\left(\overline{E}|A\right)P(B)P(C) + P(E|A)\right) \\ &= (1-\beta)q_A(t)q_B(t)q_C(t) + \beta q_A(t) \\ &= q_A(t)q_B(t)q_C(t) + \beta q_A(t)(1 - q_B(t)q_C(t)) \\ &= q_A(t)q_B(t)q_C(t) + q_E(t)(1 - q_B(t)q_C(t)), \end{aligned} \tag{71}$$

and the reliability function of the system is obtained,

$$R_{1-out-of-3}(t) = 1 - P(SF)(t). \tag{72}$$

Thus the Birnbaum reliability importance is deduced from the probability importance,

$$\begin{aligned} RIB_i(t) &= \frac{\partial R_{1-out-of-3}(t)}{\partial R_i(t)} = \frac{\partial P(SF)(t)}{\partial q_i(t)} \\ &= \begin{cases} (1-\beta)q_B(t)q_C(t) + \beta, & i = A \\ (1-\beta)q_A(t)q_C(t), & i = B \\ (1-\beta)q_A(t)q_B(t), & i = C. \end{cases} \end{aligned} \tag{72}$$

6.3.2 CTMC Model

Figure 10 illustrates the CTMC of 1-out-of-3 system. In this figure, the white and gray nodes represent operational and failed states, respectively. The state notations based on the current conditions of components are shown in Table 17. We consider three concurrent failures defined as states D1, D2, and D3. Note that state D4 represents the system failure caused by the independent failures of A, B, and C.

The system reliability at time $t = 50$ h is shown in Table 18. The classical method corresponds to the importance analysis using structure function. Both classical and proposed methods achieve the same results. Table 19 shows the time-dependent failure rate and the reliability of each component at time $t = 50$ h. It is observed that the reliabilities of components B and C are lower than that of component A due to their higher failure rates. Besides, the occurrence probability of each system failure

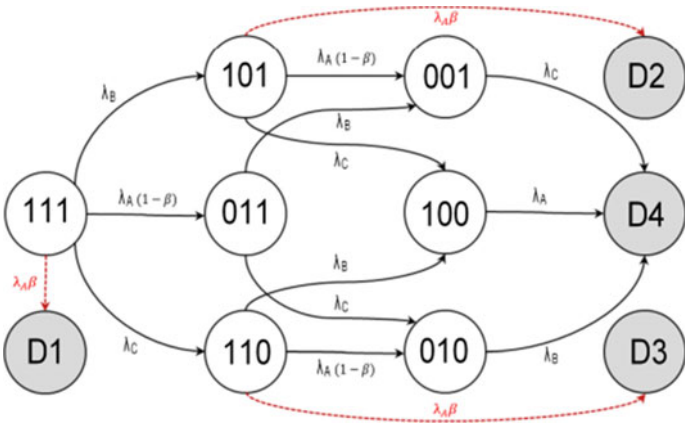


Fig. 10 CTMC of 1-out-of-3 system

Table 17 The states of 1-out-of-3 system

State	Description
111	All components are operational
101	B is failed, A and C are operational
011	A is failed, B and C are operational
110	C is failed, A and B are operational
001	A and B are failed, C is operational
100	B and C are failed, A is operational
010	A and C are failed, B is operational
D1	The concurrent failure of A, B, C
D2	The concurrent failure of A, C
D3	The concurrent failure of A, B
D4	All components are failed independently

event is given in Table 20. The probability that all components fail independently is the highest, followed by the probability that all components fail simultaneously. The occurrence probabilities of concurrent failures of, A and C, and, A and B, are the same.

Table 21 compares the classical and proposed methods in terms of Birnbaum reliability component importance at time $t = 50$ h. Although the importance values under the two methods are somewhat different, their rankings are consistent. The other importance measures [i.e., $RICR(50)$ and $RIU_{\lambda}(50)$] and their corresponding importance rankings obtained by the proposed approach are presented in Table

Table 18 System reliability ($t = 50$ h)

Method	Reliability
Classical method	0.999787
Proposed method	0.999787

Table 19 Time-dependent failure rate and reliability of components ($t = 50$ h)

Component	Failure rate	Reliability
A	0.001000	0.951229
B	0.001002	0.951137
C	0.001002	0.951137

Table 20 Occurrence probability of each failure event of system

Event	Occurrence probability
D1	9.286e−5
D2	2.301e−6
D3	2.301e−6
D4	1.158e−4

Table 21 Comparison of classical and proposed methods regarding Birnbaum reliability importance ($RIB(t)$) of components ($t = 50$ h)

Component	Classical method	Ranking	Proposed method	Ranking
A	4.374e−3	1	4.365e−3	1
B	2.374e−3	2	2.375e−3	2
C	2.374e−3	2	2.375e−3	2

Table 22 Criticality importance and upgrading function measures of components in the 1-out-of-3 system obtained using the proposed method ($t = 50$ h)

Component	$RICR(50)$	Ranking	$RIU_{\lambda}(50)$	Ranking
A	4.153e−3	1	−2.076e−4	1
B	2.260e−3	2	−1.132e−4	2
C	2.260e−3	2	−1.132e−4	2

22. The table shows that the importance rankings of components among different measures are the same.

7 Summary

In this paper, we presented the Markov chain-based component-wise sensitivity analysis approach to evaluate the component importance measures without any system structure function. Concretely, we introduced both transient and stationary analysis of Markov reward models and then presented the parametric sensitivity of MRMs. Furthermore, the component-wise sensitivity analysis based on the parametric sensitivity of MRMs and the component importance analysis using component-wise sensitivity were proposed. Besides, three types of component importance measures were derived from the viewpoints of both steady-state availability and reliability. In the numerical examples, we validated the proposed approach with three systems and showed that the proposed method did work well. In the future, we will improve and extend the method to solve more complicated models, such as non-Markovian models like Markov regenerative processes.

References

1. Fricks RM, Trivedi KS (1997) Modeling failure dependencies in reliability analysis using stochastic petri nets. In: Proceedings of 11th European simulation multi-conference, Turkey ACM Press, Istanbul, pp 355–368
2. Looss B, Saltelli A (2017) Introduction to sensitivity analysis. Handbook of uncertainty quantification. Springer, Cham, pp 1103–1122

3. Fricks RM, Trivedi KS (2003) Importance analysis with Markov chains. In: Annual reliability and maintainability symposium, 89–95
4. Zheng J, Okamura H, Dohi T (2015) Availability importance measures for virtualized system with live migration. Appl Math 6(2):359–372
5. Birnbaum ZW (1969) On The importance of different components in a multicomponent system. In: Krishnaiah PR (ed) Multivariate analysis. Academic Press, New York, pp 581–592
6. Lambert HE (1975) Measure of importance of events and cut sets in fault trees. In: Barlow RE, Fussell JB, Singpurwalla ND (eds) Reliability and fault tree analysis. SIAM, Philadelphia, PA, USA, pp 77–100
7. Beeson S, Andrews JD (2003) Importance measures for non-coherent-system analysis. IEEE Trans Reliab 52(3):301–310
8. Shen JY, Cui LR, Du SJ (2015) Birnbaum importance for linear consecutive–out-of- systems with sparse. IEEE Trans Reliab 64:359–375
9. Si S, Dui H, Zhao X, Zhang S, Sun S (2012) Integrated importance measure of component states based on loss of system performance. IEEE Trans Reliab 61(1):192–202
10. Dui H, Si S, Zuo MJ, Sun S (2015) Semi-Markov process-based integrated importance measures for multi-state systems. IEEE Trans Reliab 64(2):754–765
11. Kristiansen M, Winther R, Natvig B (2010) On component dependencies in compound software. Int J Reliab Qual Saf Eng 17(5):464–493
12. Pan ZJ, Nonaka Y (1995) Importance analysis for the systems with common cause failures. Reliab Eng Syst Saf 50:297–300
13. Reibman A, Trivedi KS (1988) Numerical transient analysis of Markov models. Comput Oper Res 15(1):19–36
14. Trivedi KS, Bobbio A (2017) Reliability and availability engineering: modeling, analysis, and applications. Cambridge University Press
15. Rutishauser H (1968) Once again: the least square problem. Linear Algebra Appl 1(4):479–488
16. Henley EJ, Kumamoto H (1981) reliability engineering and risk assessment. Prentice-Hall, Englewood Cliffs
17. Cassady CR, Pohl EA, Jin S (2004) Managing availability improvement efforts with importance measures and optimization. IMA J Manag Math 15:161–174
18. Trivedi KS (2001) Probability and statistics with reliability, queueing, and computer sciences applications, 2nd edn. Wiley, New York
19. Lanus M, Yin L, Trivedi KS (2003) Hierarchical composition and aggregation of state-based availability and performability models. IEEE Trans Reliab 52:44–52
20. Heising CD, Guey CN (1984) A comparison of methods for calculating system unavailability due to common cause failures: the beta factor and multiple dependent failure fraction methods. Reliab Eng 8(2):101–116
21. Plateau B, Stewart WJ (2000) Stochastic automata networks. In: Grassmann WK (eds) Computational probability. Kluwer Academic, University of Saskatchewan, Boston

Failure Rate Modeling of Mechanical Components and Systems

Liyang Xie

Abstract For products subjected to many times of load action during service, product life is dominated by load and its capability against load, referred to as strength. This chapter introduces load-strength interference analysis based failure rate modelling method, develops component and system failure rate models, and illustrates the causal relation between failure rate curve shape and load/strength characteristics. For the majority of mechanical components and systems, service load can be described as a random process, material property degrades during load actions, and the dynamic load-strength relationship makes the failure rate change continuously. As failure occurs on load exceeding strength, failure rate models are developed by analyzing the competition behavior between load and strength. By such failure rate models, the effects of load uncertainty, strength uncertainty and strength degradation pattern on failure rate curve shape are demonstrated. Meanwhile, the three stages of the bathtub curve are interpreted in terms of stochastic load-strength competition behavior, the roller coaster type failure rate curve is attributed to the strength diversity of the products in a population.

Keywords Failure rate definition · Life distribution · Strength degradation · Load-strength interference · Bathtub curve

1 Introduction

Failure rate is a frequently used metric for product reliability. By definition, failure rate at time t is the limit of the probability that a product will fail in a time interval $(t, t + \Delta t]$ when Δt approaches to zero, given the product is functioning at time t. Besides direct estimation based on product life data, failure rate function can be derived from the probability density function of product life by the formula $\lambda(t) = f(t)/R(t)$, i.e., failure rate at time t equals to the ratio of the life probability density at time t to the reliability over time t. This formula presents a one to one mapping

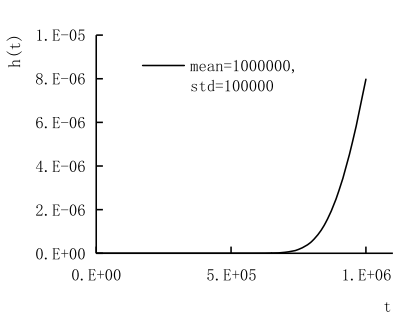
L. Xie (✉)
Department of Mechanical Engineering, Northeastern University, Shenyang 110819, China
e-mail: lyxie@mail.neu.edu.cn

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_6

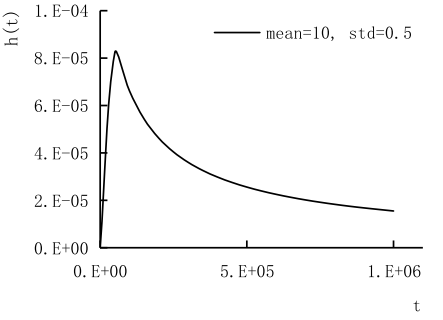
between failure rate and life distribution. It is easy to know that the exponential life distribution yields constant failure rate, the normal (Gaussian) life distribution yields increasing failure rate, the log-normal life distribution yields unimodal failure rate (first increasing and then decreasing), and the Weibull life distribution may yield increasing, decreasing or constant failure rate depending on shape parameter value (shown in Fig. 1).

On the other hand, it is traditionally believed that the failure rate curve of bathtub shape is the most typical (shown in Fig. 2). Apparently, none of the commonly used life distributions, such as the exponential distribution, the normal distribution, the log-normal distribution or the Weibull distribution can yield a failure rate curve of bathtub shape, illustrating that either the commonly used probability density functions cannot exactly describe product life distribution, or product failure rate curve does not present bathtub shape.

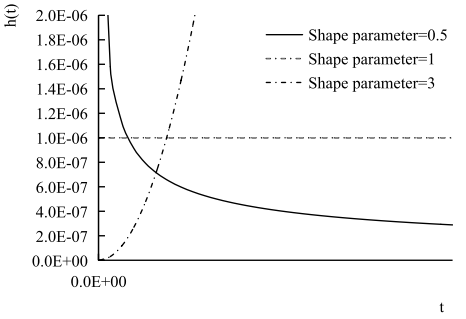
The three stages in a failure rate curve of bathtub shape as shown in Fig. 2 were conventionally partitioned as infant mortality stage (the decreasing failure rate stage appeared in the early part of the population service life), chance failure stage (the



(a) Failure rate derived from normal life distribution



(b) Failure rate derived from log-normal life distribution



(c) Failure rates derived from Weibull life distributions

Fig. 1 Failure rate curves derived from different types of life distributions

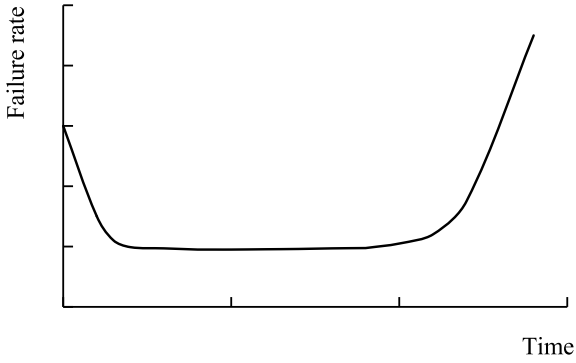


Fig. 2 A failure rate curve of bathtub shape

middle part of the failure rate curve, showing a roughly constant failure rate), and wear out stage (the increasing failure rate stage appeared in the last part of the population service life) [1]. They are also called as burn-in period, useful life period and wear-out period, respectively [2]. It is usually explained that the infant mortality stage demonstrates a sub-population dominated by quality-control defects due to poor workmanship, out-of-specification incoming parts and materials, and other substandard manufacturing practices. The other two stages were attributed to stochastic load and product performance deterioration, respectively [1]. In other words, the high failure rate in the initial phase is explained as that there are undiscovered defects in the products. These soon show up when the products are activated. When the product has survived the infant mortality period, the failure rate often stabilizes at a level where it remains for a certain amount of time until it starts to increase as the products begin to wear out [2].

The features of product failure rate have been analyzed from the aspects of reliability function [3], life distribution [4–8] and strength degradation [9, 10]. Some studies on failure rate curve shape thought that mechanical products may not appear to have an infant mortality period or chance failure period [1].

In practice, product failure rates are estimated by means of various methods and models according to life data and/or censored life data. On the other hand, traditional reliability calculation is sometimes carried out based on failure rate function [11]. It means that failure rate should be obtained based on pertinent information different from life distribution or life data. To estimate failure rate directly from product life data needs a large size sample. To derive failure rate equation from product life distribution needs exact life probability density function that is hard to obtain. Therefore, modeling product failure rate in a way different from life data-based approaches is of great significance. Besides, it is helpful to get insight into the meaning of the failure rate curves of different shape.

The complex shape of a bathtub curve implies that failure rate modelling might be difficult. To develop a failure rate model, the basic influence factors must be identified first. Generally, the service time dependent variation of product failure rate depends on load characteristics, product strength, failure mechanism and other operational profile [12–14].

For mechanical components, it is well known that load-strength interference analysis is the most widely applied method to develop reliability model [15]. However, not many studies have been conducted to developed load-strength interference relationship-based failure rate model.

As to the load characteristics and failure mechanisms typical for mechanical components and structures, such as deformation or fracture under static load or fatigue under cyclic load, the times of load action is a more direct parameter to characterize product service life. For instance, taking into account the effect of multiple actions of a random load, a loading number dependent failure probability formula for static strength failure (no strength degradation during load actions) was proposed [16]:

$$P(n) = 1 - \int_0^{\infty} f(x)[\int_0^{x} g(y)dy]^n dx \tag{1}$$

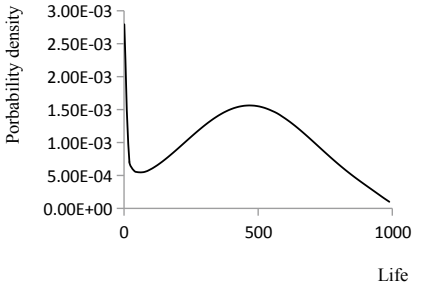
where, *P(n)* stands for the component failure probability after *n* times of load (stress) action, *f(x)* stands for the probability density function of component strength, and *g(y)* stands for the probability density function of the stress subjected to the component.

Obviously, in the situation of one time of load action, Eq. 1 degenerates into the traditional load-strength interference model for failure probability calculation:

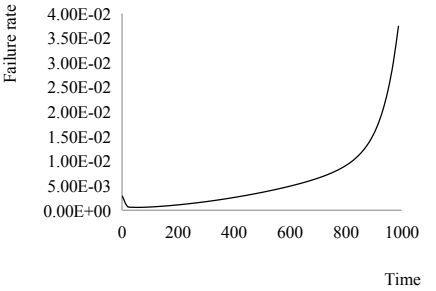
$$P = 1 - \int_0^{\infty} f(x) \int_0^{x} g(y)dydx \tag{2}$$

In principle, failure rate modeling is much the same as failure probability modeling, both can be achieved through load-strength interference analysis, since both the failure rate and the failure probability is determined by the load distribution, strength distribution, the times of load action, and the strength degradation behavior.

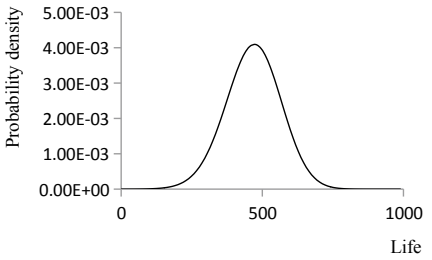
Indeed, life distribution can also be derived by means of load-strength interference analysis [17]. Different load distributions and/or strength distributions, together with their competition relations, yield different life distributions and different failure rate curves [17, 18]. Shown in Fig. 3 are the life distributions and failure rate curves of a mechanical component subjected to random loads, with strength degrading linearly during load actions. All the curves are drawn according to the respective functions formulated based on multiple variates stress-strength interference relationship. The stresses are presumed to follow the Weibull distribution, and strengths are presumed to follow the normal distribution. Figure 3a and b are the life distribution and failure rate curve in the situation of large stress dispersion and small safety margin; Fig. 3c and d are the life distribution and failure rate curve in the situation of small stress dispersion and large safety margin. Both the life distributions and the failure rate curves are considerably different for the two different stress-strength combinations. For the large stress dispersion situation, the life distribution is no longer the conventional unimodal curve, the failure rate curve presents bathtub shape; for the small



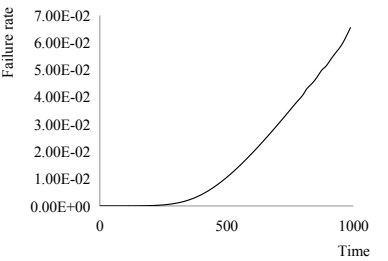
(a) Life distribution with large stress scatter



(b) Failure rate with large stress scatter



(c) Life distribution with small stress scatter



(d) Failure rate with small stress scatter

Fig. 3 Life distribution curves and failure rate curves derived from stress-strength distributions

stress dispersion situation, the life distribution presents unimodal curve, the failure rate curve is monotonically increasing.

Product life distributions are usually assumed to be one of the conventional forms such as exponential, normal, log-normal, Weibull, etc. Since none of those might be the true distribution of the product life, the failure rate function derived from life distribution may differ considerably from the true failure rate. That will mislead the understanding to the roles of the influencing factors.

This chapter introduces different ways to formulate failure rate model. First, failure rate functions are established based on stress-strength competition analysis, and the effects of stress distribution and strength distribution on failure rate curve shape are analyzed and demonstrated. Such failure rate models can clearly reveal the mechanism resulting in different shape of failure rate curves. For instance, it is illustrated that the three parts of a bathtub-shaped failure rate curve are not necessarily incurred by different root causes, different influencing factors or different failure mechanisms. Any deteriorate type of failure mechanism such as fatigue under random load history may bring about bathtub-shaped failure rate curve, whereas fatigue under constant amplitude load history leads to monotonically increasing failure rate curve. Besides, failure rate models are formulated based on the definition of failure rate in the condition that the time variable is discrete, and by virtual of random events operation.

2 Load Order Statistics and Stress-Strength Based Failure Rate Modeling

Assume that $Y_1, Y_2, \ldots, Y_n$ are independent continuous random variables with probability density function $g(y)$, $y_1, y_2, \ldots, y_n$ are n sample values of the random variables, and $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$ the sorted sample values from the minimum to the maximum. With the probability density function $g(y)$ and cumulative distribution function $G(y)$ of the random variable Y, the probability density function of the 1st order statistic (the minimum) $Y_{(1)}$ (denoted by $g_1(y)$) and that of the nth order statistic (the maximum) $Y_{(n)}$ (denoted by $g_n(y)$) are, respectively [19]:

$$g_1(y) = n[1 - G(y)]^{n-1} g(y) \tag{3}$$

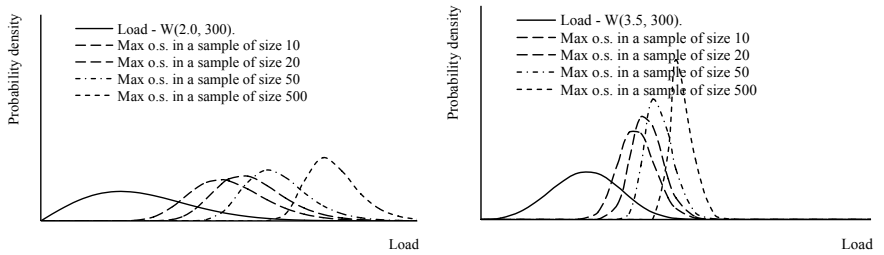
$$g_n(y) = n[G(y)]^{n-1} g(y) \tag{4}$$

For mechanical equipment and components, most of them will experience many times of random load action during service. In the situation that a product subjects to n times of random load action, the n load values can usually be treated as a set of i.i.d. (independent, identically distributed) random variables. For static strength failure, a product survives n times of load action means that product strength is greater than the maximum load (stress) appeared during the n times of load action. Therefore, the maximum statistic of the n random loads is the most direct parameter for failure rate calculation. Furthermore, it is evident that product failure probability over n times of random load action or failure rate at the nth load action depends on the times of load action as well as the scatter of the random ld variable, besides the product strength random variable.

Illustrated in Fig. 4 are the distributions of random load variables and the distributions of the corresponding maximum load order statistics in samples of size 10, 20, 50, and 500, respectively. It clearly shows that, for the random loads with different degrees of uncertainty, the distributions of the maximum load in n times of load action differ from each other considerably.

Incorporating stress order statistic into the conventional stress-strength interference model for multiple times of load action situations, dynamic (load action number dependent) component failure probability model and failure rate model can be developed. The dynamic characteristic of such models is attributed to the ever changing distribution of the maximum load order statistic. That is, the distribution of the maximum load order statistic changes continuously with the increase of the load action number. Component failure probability after n times of load action can be modeled as the following equation which is equivalent to Eq. 1 (see Fig. 5):

$$P(n) = \int_0^\infty f(x) \int_x^\infty g_n(y) dy dx \tag{5}$$



(a) Larger load dispersion (b) Small load dispersion

Fig. 4 Distributions of load random variables and their maximum order statistics (o.s.—order statistic)

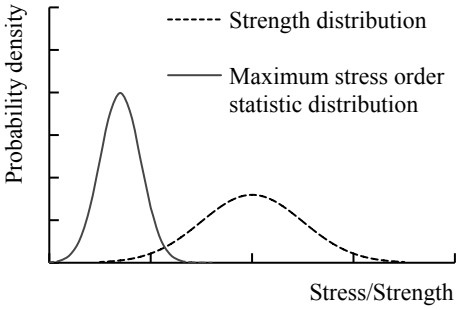


Fig. 5 Stress order statistic-strength interference relationship

In the condition that the times of load action (a discrete variable) is used as the time or life metric (it is usually treated as a continuous variable for failure rate definition), the failure rate of a product at *n*th load action, denoted by *h(n)*, can be defined as the probability of failure caused by this load action, given that the product has survived all the previous (*n*-1) times of load action. Therefore, the failure rate *h(n)* can be derived by means of the relationship between load distribution and the strength distribution of the survived products after (*n*-1) times of load action, just as product failure probability can be derived by means of the relationship between load distribution and the strength distribution of the product population. To determine failure rate *h(n)* by means of load-strength interference relationship, it is necessary to know the strength distribution of the survived products after (*n*-1) times of load action (denoted by *f(x,n)*). Obviously, the strength of a survived product will not be lower than the maximum load in the (*n*-1) times of load action (assuming no strength degradation during the (*n*-1) times of load action). Based on the strength distribution of the survived products, failure rate can be expressed as

$$h(n) = \int_0^\infty f(x,n) \int_x^\infty g(y)dydx \tag{6}$$

where, *f(x, n)* denotes the strength distribution of the survived products after (*n*-1) times of load action. Note that the strength of the product population is a random variable distributed in 0 ~∞ ;, whereas the strength of the products survived (*n*-1) times of load action is a random variable distributed in z_{n-1} ~∞ ; (z_{n-1} is the maximum stress value corresponding to the maximum load in the (*n*-1) times of load action).

Generally, the strength distribution of the survived products can be obtained by updating the original strength distribution (shown in Fig. 6). As a probability density function, it has to satisfy $\int_0^\infty f(x,n)dx = 1$. It is easy to know that the strength distribution of the products survived (*n*-1) times of load action is

$$\begin{aligned} f(x,n) &= 0 \quad \left(0 < x \le z_{(n-1)}\right) \\ f(x,n) &= \frac{f(x)}{\int_{z(n-1)}^{\infty} f(x)dx} \quad (z_{(n-1)} < x < \infty) \end{aligned} \tag{7}$$

where, $z_{(n-1)}$ denotes the maximum stress value in (*n*-1) times of load action.

As load is a random variable, the maximum stress $z_{(n-1)}$ appeared during (*n*-1) times of load action is a random variable following the maximum order statistic distribution of the stress variable. According to the total probability theorem, a failure rate model can be developed (Ref. Figure 7. For the sake of simplification, $z_{(n-1)}$ is denoted simply by z in the following). That is, failure rate *h*(*n*) equals to the statistical average of the probability that failure occurs at the *n*th load action given survived the previous (*n*-1) times of load action, weighted by the probability distribution of the maximum stress $z_{(n-1)}$ appeared during the (*n*-1) times of load action (the probability density function of $z_{(n-1)}$ is denoted by $g_{n-1}(z)$):

$$h(n) = \int_0^\infty g_{n-1}(z)\left\{\int_z^\infty f(x,n)\left[\int_x^\infty g(y)dy\right]dx\right\}dz \tag{8}$$

or

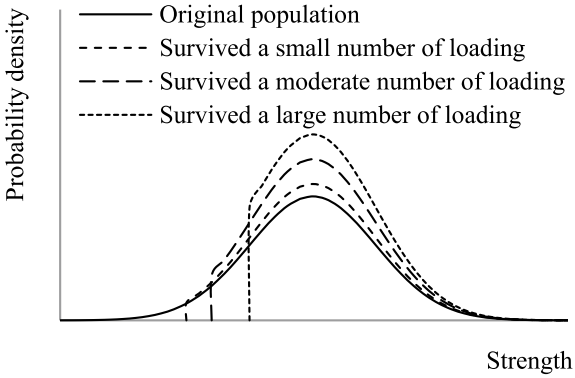


Fig. 6 Strength distribution of product population and those of the products survived a certain times of load action

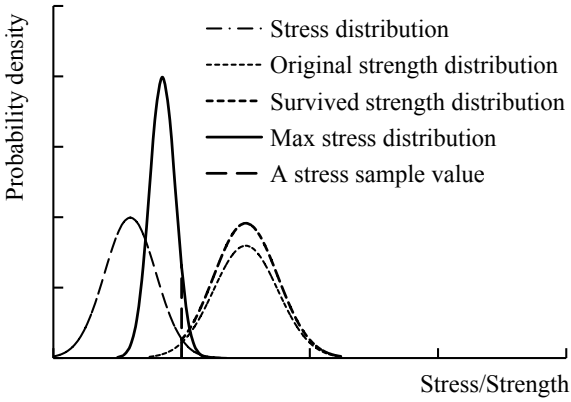


Fig. 7 Variables and their distributions involved in failure rate modeling

$$h(n) = \int_0^{\infty} g_{n-1}(z) \left\{ \int_z^{\infty} g(y) \left[\int_z^{y} f(x,n)dx \right] dy \right\} dz \tag{9}$$

In the situation of deterministic load, i.e., when a constant load is applied n times to a product, Eq. 9 degenerates as

$$h(1) = \int_0^{y} f(x)dx \tag{10}$$

$$h(n) = \int_y^{y} f(x,n)dx = 0 \quad (n \geq 2) \tag{11}$$

where, y is the constant stress value produced by the constant load.

These two equations demonstrate that, in the condition that a product subjects to many times of action of the same load, the failure rate at the first time of load action is equal to the failure probability of the product subjected to one time of load action; the failure rate at a load action number equal or greater than two equals to zero. It is easy to understand that a product survived one time of load action will survive forever, since it means that the strength of the product is greater than the stress, and it is assumed that the strength keeps the same during the load actions.

Equations 8 and 9 can also be respectively written as

$$h(n) = \int_0^{\infty} g_{n-1}(z) \left\{ \int_z^{\infty} \frac{f(x)}{1 - F(z)} \left[\int_x^{\infty} g(y)dy \right] dx \right\} dz \tag{12}$$

$$h(n) = \int_0^{\infty} g_{n-1}(z) \left\{ \int_z^{\infty} g(y) \left[\int_z^{y} \frac{f(x)}{1 - F(z)} dx \right] dy \right\} dz \tag{13}$$

and

$$h(n) = \int_0^{\infty} g_{n-1}(z) \left\{ \int_z^{\infty} g(y) \left[\frac{F(y) - F(z)}{1 - F(z)} \right] dy \right\} dz \tag{14}$$

where, F(•) stands for the cumulative distribution function of product strength.

The above failure rate equations are developed based on load-strength interference relationship. The strength distribution can be either component strength or system strength, i.e., a product can be either a component or a system. For a series system (any component failure results in system failure) composed of *m* components, denoting by $F_i(x)$ the strength distribution of component *i*, the strength distribution of such a system is

$$F_{seri}(x) = 1 - \prod_{i=1}^{m} [1 - F_i(x)] \tag{15}$$

for a parallel system (system failure occurs if and only if all components fail) composed of *m* components, the system strength distribution is

$$F_{para}(x) = \prod_{i=1}^{m} F_i(x) \tag{16}$$

In the situation that all components in a system simultaneously subject to the same load, system failure rate can be modeled as

$$h(n) = \int_0^{\infty} g_{n-1}(z) \left\{ \int_z^{\infty} g(y) \frac{F_{sys}(y) - F_{sys}(z)}{1 - F_{sys}(z)} dy \right\} dz \tag{17}$$

where, $F_{sys}(x)$ stands for either the series system strength distribution function $F_{seri}(x)$ or the parallel system strength distribution function $F_{para}(x)$.

3 Failure Rate Modeling Based on the Definition with Discrete Time Variable

Product failure rate function $h(n)$ can also be derived through failure rate definition and related events operation. Denote the event that a product fails at the *n*th load

action by A_n, denote the event that no failure occurs during the preceding (n-1) times of load action by B_{n-1}, the failure rate at nth load action is the probability that event A_n occurs given that event B_{n-1} has occurred, i.e.

$$h(n) = P(A_n|B_{n-1}) \tag{18}$$

According to the conditional probability theorem, the product failure rate (Eq. 18) can be expressed as

$$h(n) = P(A_n|B_{n-1}) = P(A_n B_{n-1})/P(B_{n-1}) \tag{19}$$

Denote the probability that product life N equals to n by $P(N = n)$, denote the product failure probability over n times of load action, i.e., the cumulative probability of product life corresponding to n times of load action by $P(n)$, i.e., $P(n) = \mathrm{P}(N \leq n)$.

It is easy to know that

$$P(A_n B_{n-1}) = P(N = n) = P(n) - P(n-1)$$

and

$$P(B_{n-1}) = 1 - P(n-1) = R(n-1)$$

Therefore,

$$\begin{aligned} h(n) &= \frac{P(n) - P(n-1)}{R(n-1)} = \frac{(1 - R(n)) - (1 - R(n-1))}{R(n-1)} \\ &= \frac{R(n-1) - R(n)}{R(n-1)} = 1 - \frac{R(n)}{R(n-1)} \end{aligned} \tag{20}$$

Equation 20 is equivalent to the failure rate definition in the situation of discrete time variable, where failure rate at the nth time of load action is defined as the probability that the product fails to the nth time of load action given functioning over the (n-1) times of load action, i.e.,

$$\begin{aligned} h(n) &= \frac{P((n-1) < N \leq n)}{P(N > (n-1))} = \frac{P(n) - P(n-1)}{R(n-1)} \\ &= \frac{R(n-1) - R(n)}{R(n-1)} = 1 - \frac{R(n)}{R(n-1)} \end{aligned} \tag{21}$$

It is equivalent to the conventional form of the failure rate defined in the situation of continuous time variable:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < N \leq (t + \Delta t))}{P(N > t) \cdot \Delta t} = \lim_{\Delta t \to 0} \frac{P(t + \Delta t) - P(t)}{R(t) \cdot \Delta t} = \frac{f(t)}{R(t)} \tag{22}$$

From Eq. 20,

$$h(n) = 1 - \frac{\int_0^\infty f(x)\left(\int_0^x g_n(y)dy\right)dx}{\int_0^\infty f(x)\left(\int_0^x g_{n-1}(y)dy\right)dx} \tag{23}$$

Or

$$h(n) = 1 - \frac{\int_0^\infty g_n(y)\left(\int_y^\infty f(x)dx\right)dy}{\int_0^\infty g_{n-1}(y)\left(\int_y^\infty f(x)dx\right)dy} \tag{24}$$

On the other hand, event A_n and B_{n-1} occur simultaneously means that failure occurs and only occurs at the nth time of load action during all the n times of loading. The probability that the event A_n (the event that product fails at the nth load action) and the event B_{n-1} (the event that no failure occurs during the preceding (n-1) times of load action) occur simultaneously is

$$P(A_n B_{n-1}) = \int_0^\infty f(x)\left\{\int_x^\infty g(y)dy\left[\int_0^x g(y)dy\right]^{n-1}\right\}dx \tag{25}$$

According to Eq. 19,

$$h(n) = \frac{\int_0^\infty f(x)\left\{\int_x^\infty g(y)dy \cdot \left[\int_0^x g(y)dy\right]^{n-1}\right\}dx}{\int_0^\infty f(x)\left\{\left[\int_0^x g(y)dy\right]^{n-1}\right\}dx} \tag{26}$$

It is easy to numerically testify that the three types of failure rate equations, i.e. Eqs. 9, 23 and 26, yield perfectly coincident failure rate curves as shown in Fig. 8.

In the situation of deterministic load, i.e., when the same load is applied many times to a product, from Eq. 24

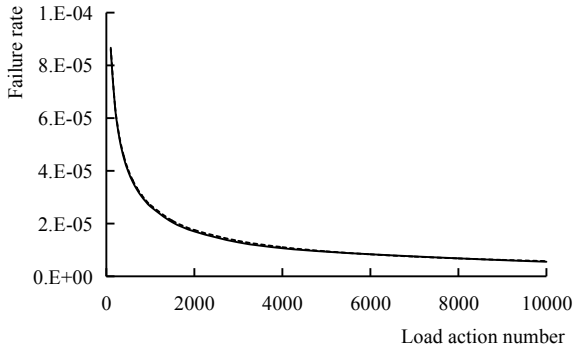


Fig. 8 Failure rate curves yielded by different equations

$$h(1) = 1 - \int_{y}^{\infty} f(x)dx = \int_{0}^{y} f(x)dx \tag{27}$$

$$h(n) = 0 \quad (n \geq 2) \tag{28}$$

from Eq. 26,

$$h(1) = \int_{0}^{\infty} f(x)\left\{\int_{x}^{\infty} g(y)dy\right\}dx = \int_{0}^{\infty} g(y)\left\{\int_{0}^{y} f(x)dx\right\}dy = \int_{0}^{y} f(x)dx \tag{29}$$

$$h(n) = 0 \quad (n \geq 2) \tag{30}$$

It illustrates that the different types of failure rate equations degenerate into the same failure rate equation in deterministic load condition.

4 Effect of Load/Strength Dispersion on Failure Rate

To demonstrate the effects of load uncertainty and strength uncertainty on product failure rate, failure rate curves corresponding to different load-strength combinations are illustrated below. Both the loads and the strengths are assumed to follow the normal distribution. The respective expectations and standard deviations are listed in Table 1. Where, μ_y stands for the mean of stress, σ_y stands for the standard deviation of stress; μ_x stands for the mean of strength, and σ_x stands for the standard deviation of strength. Failure rate curves corresponding to these four load-strength combinations are obtained by means of Eq. 9, shown in Fig. 9 as "base line", "high load std", "high strength std" and "higher load/strength std", respectively. It is demonstrated that the statistical characteristics of load distribution and strength distribution have considerable effect on failure rate curve shape.

The failure rate curves shown in Fig. 9 are decreasing because product strength is assumed no degradation during load actions. The decreasing failure rate is due to the fact that, after a certain times of load action, the survived products are those having higher strength in the population, and the products survived more times of

Table 1 Parameters of four pairs of load-strength distributions

ID	μ_y	σ_y	μ_x	σ_x
Base line	400	40	600	40
High load std	400	80	600	40
High strength std	400	40	600	80
Higher load/strength std	400	60	600	60

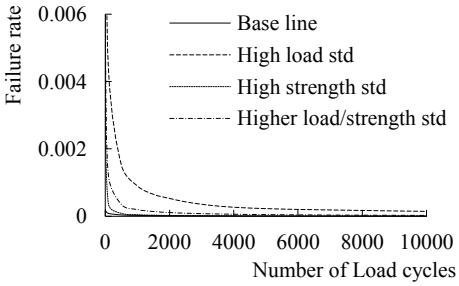


Fig. 9 Failure rate curves corresponding to different load-strength combinations

random load action have higher strength than those survived less times of random load action. Therefore, at the next time of load action, the products experienced more times of random load action have lower failure probability.

In the situation of deterministic strength, i.e., all the products have the same strength, the failure rate equation degenerates from Eq. 8 as

$$h(n) = \int_0^\infty g_{n-1}(z)\left\{\int_z^\infty f(x,n)\left[\int_x^\infty g(y)dy\right]dx\right\}dz = \int_x^\infty g(y)dy \tag{31}$$

from Eq. 23 as

$$\begin{aligned} h(n) &= 1 - \frac{\int_0^\infty f(x)\left(\int_0^x g_n(y)dy\right)dx}{\int_0^\infty f(x)\left(\int_0^x g_{n-1}(y)dy\right)dx} \\ &= 1 - \frac{\int_0^x g_n(y)dy}{\int_0^x g_{n-1}(y)dy} = 1 - \frac{G_n(x)}{G_{n-1}(x)} = 1 - G(x) \\ &= 1 - \int_0^x g(y)dy = \int_x^\infty g(y)dy \end{aligned} \tag{32}$$

from Eq. 26 as

$$\begin{aligned} h(n) &= \frac{\int_0^\infty f(x)\left\{\int_x^\infty g(y)dy \cdot \left[\int_0^x g(y)dy\right]^{n-1}\right\}dx}{\int_0^\infty f(x)\left\{\left[\int_0^x g(y)dy\right]^{n-1}\right\}dx} \\ &= \frac{\int_x^\infty g(y)dy \cdot \left[\int_0^x g(y)dy\right]^{n-1}}{\left[\int_0^x g(y)dy\right]^{n-1}} = \int_x^\infty g(y)dy \end{aligned} \tag{33}$$

By these equations, it is illustrated that product failure rate is a constant independent on the times of load action in condition of deterministic strength. Besides, it is proved once more that the three failure rate equations, i.e., Eqs. 9, 24 and 26 are the same.

5 Strength Degradation Effect on Failure Rate

Under cyclic loading, material property will degrade gradually if the stress is high enough, and the strength becomes less and less. Different products or materials have different strength degradation patterns. Some materials show an approximate linear relationship between residual strength and load action number, others show various type of non-linear relationships. First, consider the situation that material strength degrades in power law as described by Eq. 34 and shown in Fig. 10.

$$S(n) = S_0(1 - (\frac{n}{N})^e) \tag{34}$$

where, *S(n)* stands for the material strength (residual strength) after *n* times of load action, *n* stands for the number of load actions, S_0 stands for the original material strength, *N* stands for the fatigue life under the cyclic stress, *e* is a constant characterizing strength degradation rate.

With material strength degradation behavior incorporated, the basic failure rate equation (Eq. 9) becomes

$$h(n) = \int_0^\infty g_{n-1}(z) \left\{ \int_z^\infty g(y) \left[\int_z^y \frac{f_S(x,n)}{\int_z^\infty f_S(x,n)dx} dx \right] dy \right\} dz \tag{35}$$

where, $f_S(x, n)$ stands for the probability density function of the residual strength after (*n-1*) times of load action.

Shown in Fig. 11 are the strength distribution of the original products and those subjected to a certain times of random load action. Where, "original" is the strength distribution of the new products subjected to no load action, "slightly degraded", "moderately degraded" and "seriously degraded" are the products after a small, moderate and large numbers of random load actions, respectively. Correspondingly,

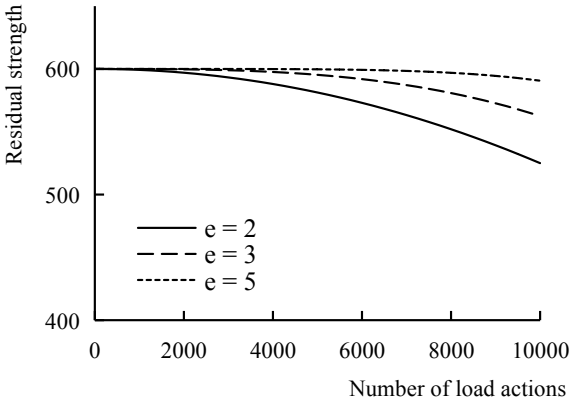


Fig. 10 Material strength degradation curves—Power law

shown in Fig. 12 are the strength distribution of the population consist of all the products subjected to no load action, and those of the sub-populations consist of products survived a small, moderate and large numbers of random load actions, respectively.

For the normal-distributed strength $S(n) \sim N(\mu_x(n), \sigma_x(n))$, suppose that the mean value of strength decreases with load action times n as

$$\mu_x(n) = \mu_0(1 - (\frac{n}{N})^e) \tag{36}$$

where, μ_0 stands for the mean of the original strength.

Under cyclic loading, the dispersion of the residual strength will also change gradually. For the sake of simplicity, especially in the situation of no enough strength degradation data available, strength standard deviation σ_x can be assumed no change, i.e.,

$$\sigma_x(n) = \text{constant} \tag{37}$$

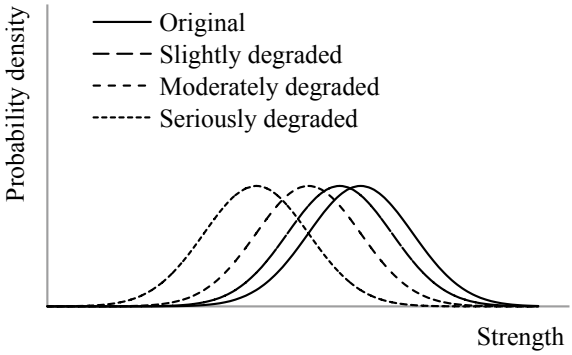


Fig. 11 Distributions of residual strength

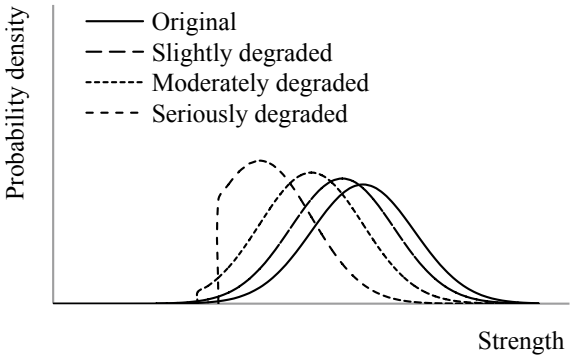
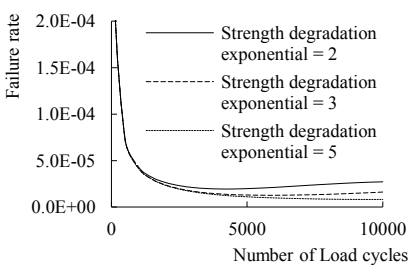
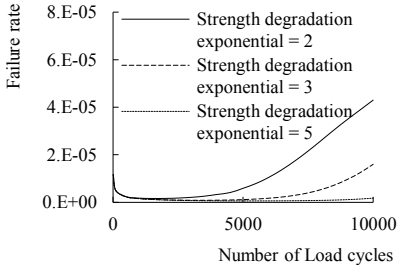


Fig. 12 Distributions of strength of survivals



(a) μ_y =400, σ_y =40; μ_x = 600, σ_x =60



(b) μ_y =450, σ_y =20; μ_x = 600, σ_x =60

Fig. 13 Failure rate curves in the situation of strength degrading in power law

The typical failure rate curves corresponding to different load-strength combinations and different strength degradation rates are obtained by Eq. 35 and shown in Fig. 13. It demonstrates that different strength degradation rates yield obviously different failure rate curves.

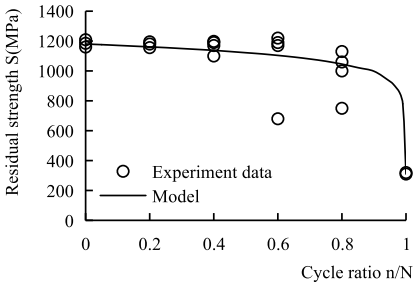
Some materials show logarithmic strength degradation described by the following equation:

$$S(n) = S_0 + (S_0 - \sigma)\ln(1 - n/(N+1))/\ln(N+1) \tag{38}$$

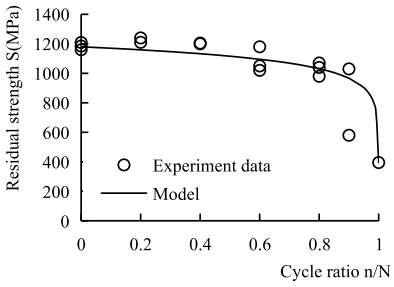
where, *S(n)* stands for residual strength after *n* times of load action, *n* stands for the number of load actions, *N* stands for the number of load action to material fatigue failure, i.e., the fatigue life under the cyclic stress σ.

Shown in Fig. 14 are test data of specimens made of normalized carbon steel and the logarithmic strength degradation curves, in which the mean of the ultimate tensile strength, i.e., the mean original strength is 1180 MPa.

When the logarithmic strength degradation equation is incorporated into the basic failure rate model, the typical failure rate curve of bathtub shape can be yielded by



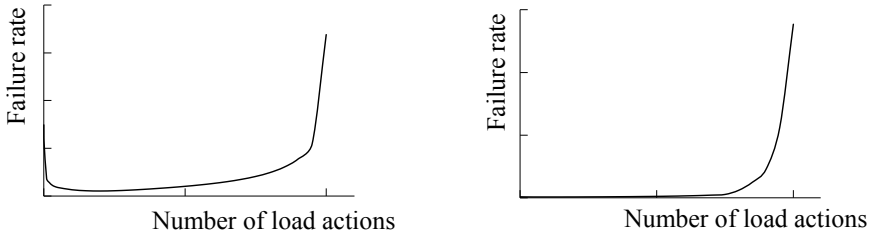
(a) Residual strength at stress 310MPa



(b) Residual strength at stress 395MPa

Fig. 14 Strength degradation test data and the logarithmic residual strength curves

Eq. 35 (shown in Fig. 15a). The typical bathtub curve is for a component in the condition that stress follows the normal distribution N(450, 20^2), strength follows the normal distribution $N\left(600+300\times\frac{\ln(1-n/(N+1))}{\ln(N+1)}, 30^2\right)$. Comparing with the failure rate curves obtained in the situation of strength degrading in power law, it demonstrates that strength degradation pattern influences the shape of failure rate curve considerably. Besides, both load distribution and strength distribution influence failure rate curve shape considerably, too. In the extreme situation of deterministic load, failure rate will increase monotonically with the degradation of strength, so it is not surprised to get a monotonically increasing failure rate curve as shown in Fig. 15b. The different curves in Fig. 16 correspond to different load distribution—strength distribution combinations. Shown in Fig. 17a are the failure rate curve of a component in the condition that stress follows the normal distribution N(450, 20^2), strength follows the normal distribution $N\left(600+300\times\frac{\ln(1-n/(N+1))}{\ln(N+1)}, 40^2\right)$, and the failure rate curve of a series system composed of 10 identical components, Fig. 17b are a failure rate curve of a component and that of a parallel system consist of 10 components.



(a) Bathtub shaped failure rate curve (b) Failure rate under deterministic load

Fig. 15 Failure rate curves yielded by Eq. 35 in the situation of logarithmic strength degradation

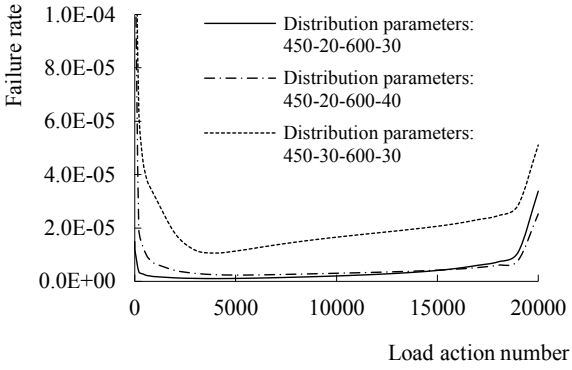
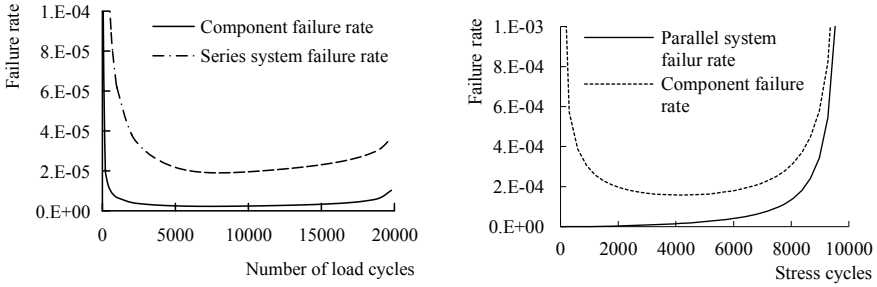


Fig. 16 Failure rate curves corresponding to different load/strength distributions (*note*: parameters shown on graph: stress mean-stress std-strength mean-strength std)



(a) Component and series system (b) Component and parallel system

Fig. 17 Failure rate curves of a components and systems consist of 10 components

The above situations demonstrate that products (components or systems) failure rate curves present on different shapes depending on system configuration, the underlying failure mechanism and stress-strength relationship of the components.

6 Mechanism to Yield Failure Rate Curve of Roller Coaster Shape

As mentioned above, the shape of product failure rate curve depends on many factors including product property, load characteristics, failure mechanism, etc. Product failure rate may be as simple as a constant value, a monotonically increasing curve or a monotonically decreasing curve. It may also be as complicated as a bathtub curve or a roller coaster curve as shown in Fig. 18.

The roller-coaster type of failure rate curve shown in Fig. 18a is obtained in the situation that the product population can be divided into two sub-populations. Two potential failure modes exist for the population. A product in the population might fail in either failure mode with a respective probability. Here, one failure mode is related to Weibull distributed stress with shape parameter 2.0 and scale parameter 200,

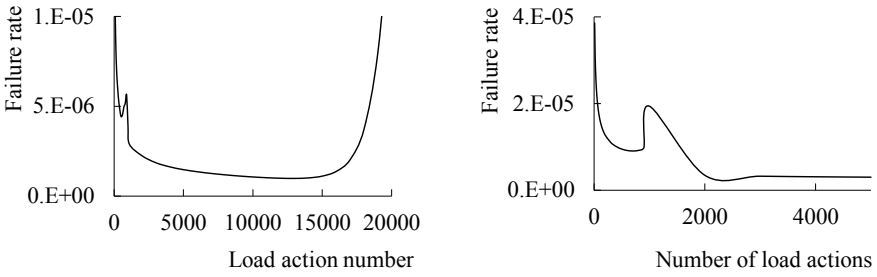


Fig. 18 Failure rate curves of roller coaster shape

normal distributed strength with original mean 600 and standard deviation 60, the mean strength degrades in logarithmic law with a life index 2000; the other is related to normal distributed strength with original mean 600 and standard deviation 120, the strength mean degrades in logarithmic law with a life index 1000. The probability for the first failure mode to take place is 0.7, and that for the second failure mode is 0.3. That is, 70% products in the population belong to one sub-population, the other 30% belong to another sub-population.

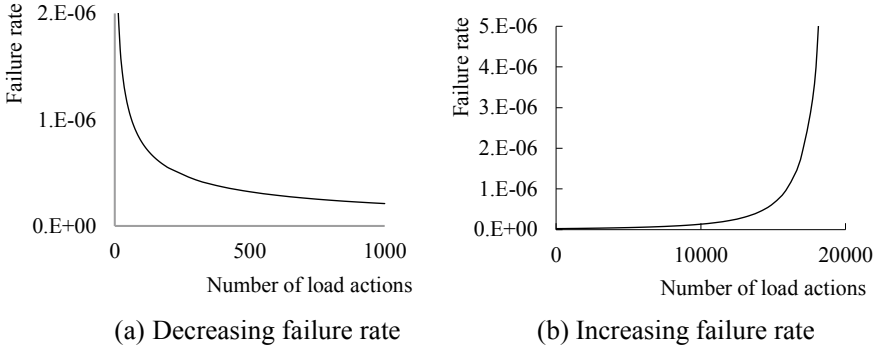
Shown in Fig. 18b is the situation of Weibull distributed stress with shape parameter 2.0 and scale parameter 150, strength characteristics are the same as mentioned above despite that the probability for the first failure mode to take place is 0.9, and that for the second failure mode is 0.1.

By the way, for a failure rate curve of bathtub shape, the high failure rate in the infant mortality phase is traditionally attributed to products with material flaw or manufacture defect. The fact is that, if the products can be divided into two groups as perfect and defective, the failure rate curve will present roller coaster shape as shown in Fig. 18. In such a product population, part of the products has manufacture defects manifested as lower strength and higher failure probability in the infant motility stage, others do not have manufacture defect, and therefore, their failures occur mainly in wear out stage.

Furthermore, it is traditionally believed that after the failures of the defective products, the survived products will keep a low and roughly constant failure rate for a long time until they begin to wear out. The failure, i.e., the so-called chance failure in this period is attributed to unexpected factors. Such an explanation is also plausible, since if there are the so called “unexpected factors”, they will exist throughout the product service life, not only in the chance failure stage. In fact, the failure rate model has illustrated that the roughly constant failure rate in the useful life period is the result of stress-stress competition, instead of unexpected factor.

Since product failure or not during service is determined by the dynamic competition relationship between load and strength, the decreasing failure rate in the first stage of product service life is natural for products subjected to multiple times of random load action. If a sub-population, dominated by quality-control defects due to poor workmanship, out-of-specification incoming parts and materials, and other substandard manufacturing practices, exists in the product population, there will be more than one failure mechanisms or failure modes, the corresponding failure rate curve will present roller coaster shape. In other words, if only a part of the products in a population suffers from a certain failure mechanism leading to shorter lifetime, the failure rate curve might present roller coaster shape.

Strength degradation under repeatedly loading results in increasing failure rate. Provide that product property keeps the same during its service life, i.e., no strength degrading during load actions, a product will never fail to a load not greater than those experienced. That is, such a product can only fail when a load higher than all the preceding ones is applied. For a steady random load process, the probability for a higher load to appear is less and less with the increase of the loading history. Therefore, given that strength keeps no change, failure rate will decrease with the increase of product service time as show in Fig. 19a. In the condition that strength



(a) Decreasing failure rate (b) Increasing failure rate

Fig. 19 Failure rate curves in different situations

degrades gradually, the failure rate is increasing for a component with large safety margin or a component subjected to a deterministic load as shown in Fig. 19b.

Although the shapes of failure rate curve are various, failure is essentially the result of the competition between load and strength. In this regard, whether a product fails at a certain load action or not depends on the load and the strength at that moment. Likewise, failure rate at a certain load action number is determined by load distribution and strength distribution including its time dependent degradation behavior. Based on load-strength competition behavior, it is easy to understand that the failure rate curve of bathtub shape comes from the continuously changing competition relationship between the random load and degrading strength, the failure rate curve of roller coaster shape manifests the diversity of product strength in the same population.

7 Conclusion

Failure rate models are developed for components and systems based on dynamic load-strength competition analysis. It is illustrated that whether the failure rate, as a function of the times of load action, takes on bathtub shape or not is mainly determined by stress-strength relationship. The reason for decreasing failure rate is the less and less probability that a higher load appears with the increase of preceding load action numbers, the increasing failure rate is caused by strength degradation. The models, together with the failure rate curves corresponding to the models, highlight the effect of the statistical characteristics of load and strength on the shape of failure rate curve, as well as the role of strength degradation.

It is clearly demonstrated that if product strength doesn't degrade, failure rate curve takes on the feature of the first two stages of a typical three-stage bathtub-shaped curve only, i.e., failure rate decreases continuously with the number of load actions, with lower and lower gradient. When the effect of strength degradation exceeds the effect of the decreasing probability of higher load appearing, the failure

rate begins to increase. In other words, the decreasing failure rate stage is dominated by the statistical risk of load, whereas the increasing failure rate stage is dominated by strength degradation.

For a population containing defective products, the failure rate curve may present roller coaster shape. Generally, a failure rate curve of roller coaster shape will appear in a population where different failure mechanisms or different failure modes exist in different products.

References

1. Wasserman G (2002) Reliability verification, testing, and analysis in engineering design. CRC Press
2. Rausand M, Hoyland A (2004) System reliability theory, 2nd edn. Wiley, Hoboken
3. Wang KS, Hsu FS, Liu PP (2002) Modeling the bathtub shape hazard rate function in terms of reliability. Reliab Eng Syst Saf 75(3):397–406
4. Ghai GL, Mi J (1999) Mean residual life and its association with failure rate. IEEE Trans Reliab 48(3):262–266
5. Tang LC, Lu Y, Chew EP (1999) Mean residual life of lifetime distributions. IEEE Trans Reliab 48(1):73–78
6. Bebbington M, Lai CD, Zitikis R (2007) A flexible Weibull extension. Reliab Eng Syst Saf 92(6):719–726
7. Moan T, Ayala-Uraga E (2008) Reliability-based assessment of deteriorating ship structures operating in multiple sea loading climates. Reliab Eng Syst Saf 93(3):433–446
8. Xie M, Lai CD (1996) Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. Reliab Eng Syst Saf 52(1):87–93
9. Bae SJ, Kuo W, Kvam PH (2007) Degradation models and implied lifetime distributions. Reliab Eng Syst Saf 92(5):601–608
10. Meeker WQ, Escobar LA, Pascual FG (2022) Statistical methods for reliability data. Wiley
11. Li W, Pham H (2005) Reliability modeling of multi-state degraded systems with multi-competing failures and random shocks. IEEE Trans Reliab 54(2):297–303
12. Xie L, Zhou J, Wang Y, Wang X (2005) Load-strength order statistics interference models for system reliability evaluation. Int J Performability Eng 1(1):23
13. Abunima H, Teh J (2020) Reliability modeling of PV systems based on time-varying failure rates. IEEE Access 8:14367–14376
14. Jiang R (2013) A new bathtub curve model with a finite support. Reliab Eng Syst Saf 119:44–51
15. Haines DJ (2012) Practical reliability engineering–Fifth edition, PDT O'Connor and A. Kleyner, Wiley, West Sussex UK
16. Xie L, Wang Z, Lin W (2008) System fatigue reliability modelling under stochastic cyclic load. Int J Reliab Saf 2(4):357–367
17. Xie L, Wang Z (2009) Load-strength interference failure rate model. Int J Reliab Qual Saf Eng 16(03):249–260
18. Xie LY, Qin B (2017) In: Pham H (ed) Proceeding of reliability and quality in design. Chicago, USA
19. Larsen RJ, Marx ML (2001) An introduction to mathematical statistics and its application. Prentice Hall, Boston

Statistical Reliability Modeling and Analysis for Repairable Systems with Bathtub-Shaped Intensity Function

Suk Joo Bae, Byeong Min Mun, and Paul H. Kvam

Abstract The nonhomogeneous Poisson process (NHPP) has become a useful approach for modeling failure patterns of recurrent failure data revealed by minimal repairs from an individual repairable system. This work investigates complex repairable artillery systems that include several failure modes. We propose a superposed log-linear process (S-LLP) based on a mixture of nonhomogeneous Poisson processes in a minimal repair model. This allows for a bathtub-shaped failure intensity that models artillery data better than currently used methods. The method of maximum likelihood is used to estimate model parameters and construct confidence intervals for the cumulative intensity of the S-LLP. Additionally, for multiple repairable systems presenting system-to-system variability, we apply the mixed-effects models to recurrent failure data with bathtub-shaped failure intensity, based on the superposed Poisson process models including S-LLP. The mixed-effects models explicitly involve between-system variation through random-effects, along with a common baseline for all the systems through fixed-effects. Details on estimation of the parameters of the mixed-effects superposed Poisson process models and construction of their confidence intervals are examined in this work. An applicative example of multiple artillery systems shows prominent proof of the mixed-effects superposed Poisson process models for the purpose of reliability analysis.

Keywords Minimal repair · Mixed-effects model · Nonhomogeneous poisson process · Power law process · Log linear process · Superposed process

S. J. Bae (✉)
Department of Industrial Engineering, Hanyang University, Seoul, Korea
e-mail: sjbae@hanyang.ac.kr

B. M. Mun
Samsung Display Co. Ltd, Asan, Korea

P. H. Kvam
Department of Mathematics and Computer Science, University of Richmond, Richmond, VA 23173, USA

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_7

1 Introduction

Modern systems consist of numerous parts working together, making the maintenance action for the systems more difficult. In general, systems can be classified into repairable and non-repairable systems according to feasibility of maintenance activity. A repairable system is one that can be restored to an operating condition without replacement of the entire system after some repair activity is executed. For the repairable system, the patterns of failures collected after successive repairs are very important to establish an effective maintenance policy. For example, increasing time intervals between failures suggest reliability improvement, while decreasing time intervals imply reliability deterioration. Repair processes of this type can emulate a minimal repair model in which the repair or the substitution of a failed part tends to have a negligible effect on overall system reliability, restoring the system performance to the exact same condition as it was just before the failure. Because the system is restored to its current state (immediately preceding the most recent failure), the assumption of minimal repair reveals a failure pattern governed by a nonhomogeneous Poisson process (NHPP). The NHPP has garnered significant attention in the reliability literature [1, 2].

2 Nonhomogeneous Poisson Process Model

The NHPP is defined by its nonnegative intensity function $\lambda(t)$. The expected number of failures in the time interval $(0, t]$ is obtained by $\Lambda(t) = \int_0^t \lambda(u)\, \mathrm{d}u$. The intensity function $\lambda(t)$ is equal to the rate of occurrence of failures (ROCOF) associated with the repairable system [2]. When the intensity function is constant, i.e., $\lambda(t) \equiv \lambda$, the process reduces to a homogeneous Poisson process (HPP). The NHPP has been widely used in modeling failure frequency for repairable systems because of its flexibility and mathematical tractability via its intensity function $\lambda(t)$ [3].

2.1 Monotonic Failure Intensity Model

The most commonly applied form of NHPP is the power law process (PLP). Crow [4] suggested a PLP model under "find it and fix it" conditions with the intensity function

$$\lambda(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}, \quad t > 0 \tag{1}$$

where β (> 0) and α (> 0) are the shape and scale parameters, respectively. The corresponding mean cumulative number of failures over $(0, t]$ is $\Lambda(t) = \left(\frac{t}{\alpha}\right)^{\beta}$. As another functional form of NHPP, a log linear process (LLP) has intensity function

$$\lambda(t) = \gamma e^{\kappa t}, \quad t > 0 \tag{2}$$

and the corresponding mean cumulative number of failures over $(0, t]$ is $\Lambda(t) = \gamma\kappa^{-1}\left(e^{\kappa t} - 1\right)$, for the parameters $\gamma(> 0)$ and κ. The LLP model was first proposed by Cox and Lewis [5] to model air conditioner failures. The PLP and the LLP models have been employed to model failure patterns of a repairable system having monotonic intensity, i.e., decreasing failure patterns (reliability improve- ment) with $\beta < 1$ ($\kappa < 0$) or increasing failure patterns (reliability deterioration) with $\beta > 1(\kappa > 0)$. When $\beta = 1(\kappa = 0)$, , the PLP (LLP) reduces to the HPP.

The intensity function of the PLP model tends to infinity as the system age increases, whereas the observed failure process may have a finitely bounded intensity function. Considering NHPPs with a finite and bounded intensity function, Pulcini [6] proposed a bounded intensity process (BIP) with intensity function

$$\lambda(t) = a\left[1 - e^{-\frac{t}{b}}\right]. \quad a, b > 0; t > 0 \tag{3}$$

The intensity function is increasing and bounded, approaching an asymptote of a as t tends to infinity.

2.2 Non-monotonic Failure Intensity Model

In some cases, a repairable system is subject to early (or infant mortality) failures due to the presence of assembly defects that are not screened out completely through the burn-in process, as well as wear-out failures caused by deteriorating phenomena. This causes a so-called *bathtub-shaped* failure intensity, which is typical for large and complex systems that are characterized by a number of different failure modes [7]. The PLP and the LLP models are too simplistic to accommodate this bathtub characteristic of the failure process. As an alternative, unions of several independent NHPPs called superposed Poisson processes (SPPs) have been developed to model this kind of non-monotonic failure intensity. When any subsystem failure can independently cause the system to break down, the superposed model is a natural model for the failure of the system. For an SPP based on J independent processes, let $N_j(t)$ be the number of failures in $(0, t]$ for the jth subsystem $(j = 1, 2, ..., J)$ with the intensity function $\lambda_j(t) = dE\left[N_j(t)\right]/dt$. The number of failures in $(0, t]$ for the system in the SPP is characterized by $\mathrm{N(t)} = \sum_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{N_j(t)}$. If $\mathrm{N_j(t)}$, $\mathrm{j} = 1, 2, ...,$ J are independent nonhomogeneous Poisson processes, then N(t) is also the NHPP with intensity function $\lambda(\mathrm{t}) = \sum_{\mathrm{j=1}}^{\mathrm{J}} \lambda_{\mathrm{j}}(\mathrm{t})$

SPPs have found successful application in modeling software reliability, where early detection and removal of coding errors can sometimes lead to reliability growth, e.g., the Musa-Okumoto process [8] for modeling recurrent errors in a software. Pulcini [9] proposed the superposition of two independent power law processes (called the "superposed power law process" (S-PLP)) to model the bathtub-shaped

failure pattern of a repairable system with intensity function

$$\lambda(t) = \frac{\beta_1}{\alpha_1}\left(\frac{t}{\alpha_1}\right)^{\beta_1-1} + \frac{\beta_2}{\alpha_2}\left(\frac{t}{\alpha_2}\right)^{\beta_2-1}, \quad \alpha_j, \beta_j > 0, \, j = 1, 2 \tag{4}$$

In Pulcini's model, the parameters β_1 and β_2 determine the failure patterns of a repairable system. For example, $\beta_1 < 1$ models the failure pattern of a system improving over time, while $\beta_2 > 1$ models that of a system deteriorating over time. As a result, the S-PLP with $\beta_1 < 1$ and $\beta_2 > 1$ is able to model a repairable system with the bathtub-shaped failure intensity. Yang and Kuo [10] proposed the superposition of the Musa-Okumoto process and the power law process as

$$\lambda(t) = \frac{\beta_1}{t+\alpha_1} + \alpha_2\beta_2 t^{\beta_2-1} \quad \alpha_j, \beta_j > 0, \, j = 1.2 \tag{5}$$

with corresponding mean cumulative number of failures $(0, t]$, $\Lambda(t) = \beta_1 \ln\left(1 + \frac{t}{\alpha_1}\right) + \alpha_2 t^{\beta_2}$. As Hjorth [11] pointed out, this intensity function has increasing, decreasing, and bathtub types of shapes. Later, Guida and Pulcini [12] proposed the bathtub bounded intensity process (BBIP) represented by the following superposed intensity function

$$\lambda(t) = ae^{-t/b} + \alpha\left(1 - e^{-\frac{t}{\beta}}\right), a, b, \alpha, \beta > 0, \tag{6}$$

where the first component represents a log-linear process with decreasing intensity function and the latter component is a bounded intensity process with increasing bounded intensity function [6]. Guida and Pulcini [12] showed that the BBIP is able to model the failure pattern of a repairable system subject to both early failures and deterioration phenomena, featuring a finite asymptote as the system age increases.

This work is mainly motivated by unscheduled maintenance data of artillery systems collected during exercise in the field over a fixed period of time from the Republic of Korea (ROK) army. Some of the artilleries are subject to early failures due to the presence of defective parts or assembling defects, as well as wear-out failures caused by deteriorating phenomena. This causes a non-monotonic trend in the failure data in which the intensity function initially decreases followed by a long period of constant or near constant intensity until wear-out finally occurs, at which time the intensity function begins to increase. We found that existing models including the S-PLP and the BBIP did not adequately capture the non-monotonic trend in the failure process for this field artilleries. Because of this, we propose a superposed log-linear process (S-LLP) to model ROK Army artillery system failures, and we derive the maximum likelihood estimators (MLEs) for the model parameters, along with their confidence intervals. Based on the NHPP models for a repairable system, we will go over the application of mixed-effects models for recurrent failure data from multiple repairable systems for the purpose of reliability analysis.

3 Superposed Log-Linear Process for Bathtub-Shaped Intensity

Consider a repairable system with failures observed over the time interval $(0, T]$. Suppose that the failures are subject to two different failure modes, and that each of the modes are modeled by an LLP with parameters α_j and β_j for $j = 1, 2.$. We propose a superposed log-linear process (S-LLP) with intensity function

$$\lambda(t) = \alpha_1 e^{-\beta_1 t} + \alpha_2 e^{-\beta_2 t}, \quad \alpha_1, \alpha_2, \beta_1, \beta_2 > 0, \tag{7}$$

for $t \geq 0$. A key difference between the S-LLP and previously mentioned SPPs is evident in the parameters β_1 and β_2. By limiting them to be strictly non-negative, the superposed process is a mixture of an increasing and a decreasing pair of intensity functions. Note that if $\beta_1 = \beta_2 = 0,$, the S-LLP is reduced to the homogeneous Poisson process (HPP) with a constant intensity $\lambda \equiv \alpha_1 + \alpha_2$. Unlike the S-PLP intensity function for $\beta_1 < 1$ or $\beta_2 < 1,$, the S-LLP intensity function (7) is finite at $t = 0$. The first derivative of intensity function $\lambda(t)$ with respect to t,

$$\lambda^{'(t)} = -\alpha_1 e^{-\beta_1 t} + \alpha_2 e^{\beta_2 t},$$

is equal to $\alpha_2\beta_2 - \alpha_1\beta_1$ at $t = 0$, hence $\lambda(t)$ is initially decreasing if and only if $\alpha_1\beta_1 > \alpha_2\beta_2$, and $\lambda'(t)$ is equal to 0 at $t = \tau$, where τ is given by

$$\tau = \tfrac{1}{\beta_1+\beta_2} \ln\left(\tfrac{\alpha_1\beta_1}{\alpha_2\beta_2}\right) \tag{8}$$

The point with minimum intensity (τ) lies between 0 and T if $0 \leq \ln(\alpha_1\beta_1/\alpha_2\beta_2) \leq (\beta_1 + \beta_2)T$. The second derivative of the intensity function is

$$\lambda''(\tau) = \alpha_1\beta_1^2 e^{-\beta_1\tau} + \alpha_2\beta_2^2 e^{\beta_2\tau} > 0, \tag{9}$$

and τ represents a unique time-point having minimum intensity value

$$\lambda(\tau) = \alpha_1 e^{-\beta_1\tau}\left(\tfrac{\beta_1+\beta_2}{\beta_2}\right). \tag{10}$$

That is, the intensity decreases until t $= \tau$, after which it increases from t $= \tau$ to t $= T$. Thus, the intensity function (7) reflects a bathtub behavior of sequential failures in a repairable system when the system is subject both to early failures and to wear-out failures. The expected number of failures up to t is given by

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u = \tfrac{\alpha_1}{\beta_1}\left(1 - e^{-\beta_1 t}\right) + \tfrac{\alpha_2}{\beta_2}\left(e^{\beta_2 t} - 1\right), \quad t \geq 0. \tag{11}$$

Similar to the S-PLP, it is the sum of expected number of failures caused by each failure mode, and it has an inflection point.

3.1 Maximum Likelihood Estimation

We consider the likelihood function for an NHPP with the first n failure-times, $\boldsymbol{t} \equiv (t_1 < t_2 < \cdots < t_n)$, which are observed until T. Under a failure-truncated sampling, the log-likelihood function of the S-LLP is

$$\begin{aligned}\ell(\alpha_1, \alpha_2, \beta_1, \beta_2; \boldsymbol{t}) = &\sum_{i=1}^{n} \ln\left[\alpha_1 e^{-\beta_1 t_i} + \alpha_2 e^{\beta_2 t_i}\right] \\ &- \left[\frac{\alpha_1}{\beta_1}\left(1 - e^{-\beta_1 t_n}\right) + \frac{\alpha_2}{\beta_2}\left(e^{\beta_2 t_n} - 1\right)\right],\end{aligned} \tag{12}$$

and t_n is replaced by T under a time-truncated sampling. The maximum likelihood estimators (MLEs) of the parameters $\boldsymbol{\theta} \equiv (\alpha_1, \alpha_2, \beta_1, \beta_2)^T$ can be found by solving the following likelihood equations:

$$\begin{aligned}\frac{\partial \ell}{\partial \alpha_1} &= \sum_{i=1}^{n} \frac{e^{-\beta_1 t_i}}{\alpha_1 e^{-\beta_1 t_i} + \alpha_2 e^{\beta_2 t_i}} - \frac{1}{\beta_1}\left(1 - e^{-\beta_1 t_n}\right) = 0, \\ \frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^{n} \frac{-\alpha_1 t_i e^{-\beta_1 t_i}}{\alpha_1 e^{-\beta_1 t_i} + \alpha_2 e^{\beta_2 t_i}} + \left[\frac{\alpha_1}{\beta_1^2}\left(1 - e^{-\beta_1 t_n}\right) - \frac{\alpha_1 t_n}{\beta_1} e^{-\beta_1 t_n}\right] = 0, \\ \frac{\partial \ell}{\partial \alpha_2} &= \sum_{i=1}^{n} \frac{e^{\beta_2 t_i}}{\alpha_1 e^{-\beta_1 t_i} + \alpha_2 e^{\beta_2 t_i}} - \frac{1}{\beta_2}\left(e^{\beta_2 t_n} - 1\right) = 0, \\ \frac{\partial \ell}{\partial \beta_2} &= \sum_{i=1}^{n} \frac{\alpha_2 t_i e^{\beta_2 t_i}}{\alpha_1 e^{-\beta_1 t_i} + \alpha_2 e^{\beta_2 t_i}} + \left[\frac{\alpha_2}{\beta_2^2}\left(e^{\beta_2 t_n} - 1\right) - \frac{\alpha_2 t_n}{\beta_2} e^{\beta_2 t_n}\right] = 0,\end{aligned} \tag{13}$$

Obviously, there is no closed form solution to the MLEs in (13), and these equations must be solved numerically. Even though $\ell(\alpha_1, \alpha_2, \beta_1, \beta_2; \boldsymbol{t})$ is an amalgamation of relatively well-behaved (generally concave) functions, a general search method such as Newton–Raphson is slow to work across four dimensions. In this work, we introduce a slightly more efficient numeric method based on a conditional likelihood method used by Cox and Lewis [5].

Once the MLEs of the model parameters have been obtained, the MLEs of other quantities of interest, such as the expected number of failures up to a given time, $\Lambda(t)$, as well as the probability distribution of the number of failures occurring in a future time interval $\Pr\{N(T, T + \Delta) = k\}$, can be given as

$$\hat{\Lambda}(t) = \frac{\hat{\alpha}_1}{\hat{\beta}_1}\left(1 - e^{-\hat{\beta}_1 t}\right) + \frac{\hat{\alpha}_2}{\hat{\beta}_2}\left(e^{\hat{\beta}_2 t} - 1\right), \tag{14}$$

and

$$\widehat{\Pr}\{N(T, T+\Delta) = k\} = \frac{\left[\hat{\Lambda}(T+\Delta) - \hat{\Lambda}(T)\right]^k}{k!} \cdot e^{-\left[\hat{\Lambda}(T+\Delta) - \hat{\Lambda}(T)\right]},$$

for $k = 0, 1, 2, \ldots$ and $\hat{\Lambda}(T) = n$.

We can construct confidence intervals for these and other functions based on standard errors derived from the (observed) Fisher information matrix. A large-sample approximation of estimated standard errors of the ML estimators is given by the estimated variance–covariance matrix $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ for $\hat{\boldsymbol{\theta}} \equiv \left(\hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2\right)^{\mathrm{T}}$, where $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is computed as the inverse of the estimated Fisher information matrix.

We are primarily interested in constructing confidence intervals for $\Lambda(t)$ instead of the basic parameter set $\boldsymbol{\theta}$. We approximate the standard error for $\Lambda(t)$ using properties of $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ and by using the delta method on (14). In general, for a differentiable real-valued function $g(\boldsymbol{\theta})$, the approximate standard error of $\hat{g} \equiv g\left(\hat{\boldsymbol{\theta}}\right)$ can be obtained by using the delta method as

$$\widehat{s.e.}\left(\hat{g}\right) = \sqrt{\sum_{i=1}^{4}\left(\frac{\partial g}{\partial \theta_i}|_{\hat{\boldsymbol{\theta}}}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\theta}_i\right) + \sum_{i=1}^{4}\sum_{j\neq i}^{4}\left(\frac{\partial g}{\partial \theta_i}|_{\hat{\boldsymbol{\theta}}}\right)\left(\frac{\partial g}{\partial \theta_j}|_{\hat{\boldsymbol{\theta}}}\right)\widehat{\mathrm{Cov}}\left(\hat{\theta}_i, \hat{\theta}_j\right)}, \tag{15}$$

where $(\theta_1, \theta_2, \theta_3, \theta_4) \equiv (\alpha_1, \beta_1, \alpha_2, \beta_2)$.

When the function $g(\boldsymbol{\theta})$ is invertible, the approximate standard error (15) is exactly same as that given by the estimated variance–covariance matrix relative to the log-likelihood function re-parameterized in term of g. On the other hand, when $g(\boldsymbol{\theta})$ is not invertible (as in the case of $\Lambda(t)$), the log-likelihood function cannot be re-parameterized directly, and the delta-method seems to be the only available method that does not require resampling methods [12]. The approximate $(1-\gamma)\%$ confidence interval for the function g results in either

$$\hat{g} \pm \mathrm{z}_{\gamma/2} \cdot \widehat{s.e.}\left(\hat{g}\right) \quad \text{or} \quad \hat{g}\exp\left\{\pm \mathrm{z}_{\gamma/2} \cdot \frac{\widehat{s.e.}\left(\hat{g}\right)}{\hat{g}}\right\}$$

using the normal approximation or the lognormal approximation, respectively. Although the normal assumptions (based on asymptotic properties of the MLE) are not perfectly realized for $\hat{\Lambda}(t)$ at small values of t, we do not consider transformations in this case because confidence intervals for $\Lambda(t)$ are of more interest at values of t not close to zero.

3.2 *Analysis of Artillery Repair Data*

The proposed model was applied to field repair data of eight sets of artillery systems. Each artillery system was subject to minimal repair at time of failure and all failure data for the eight artillery systems were treated as failure-truncated samples. As shown in Fig. 1, due to a number of failures observed during the early and final periods of data collection, the bathtub-shaped failure intensity potentially seems to be appropriate to describe the failure pattern of the artillery systems.

In practice, decisions concerning failure patterns have been made using graphical techniques or statistical trend tests [2]. The total time on test (TTT) plot [13] helps reveal failure patterns through curvature. A bathtub-shaped failure process can be observed in a TTT plot by an S-shaped function. For example, the first artillery data set (ID-1) consists of 62 failure-times observed until $t_{62} = 1,452$ hours and its TTT plot is contained in Fig. 2, which shows a clear indication of a bathtub-shaped intensity function for failure data of the system. It was observed that the

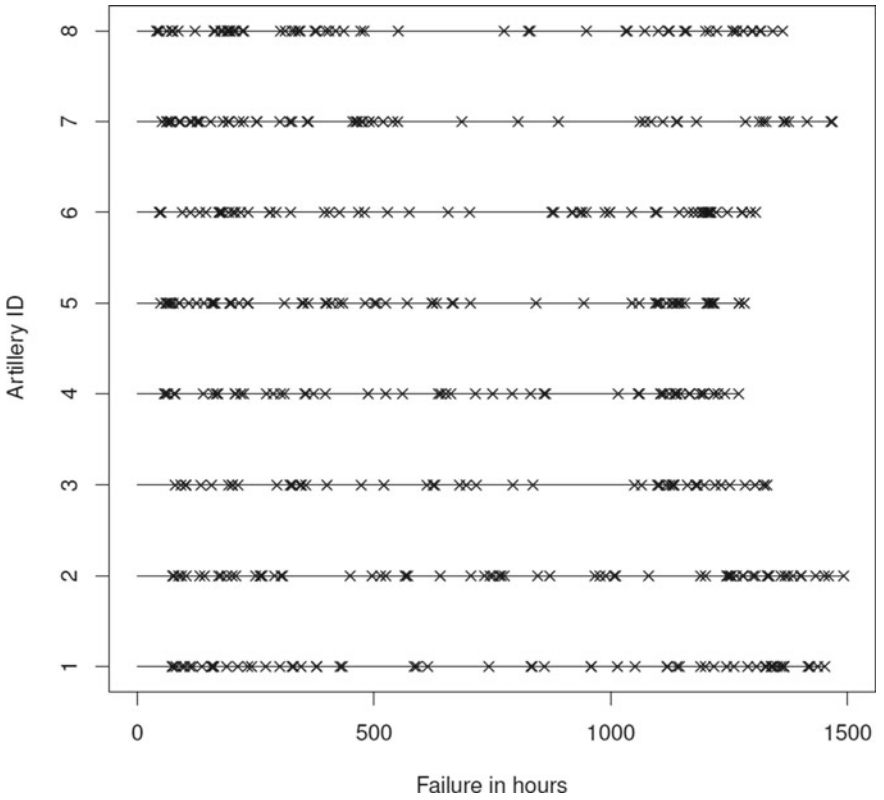


Fig. 1 Event plot showing failure times for eight sets of ROK artillery repair data across a 1500 h period of observation

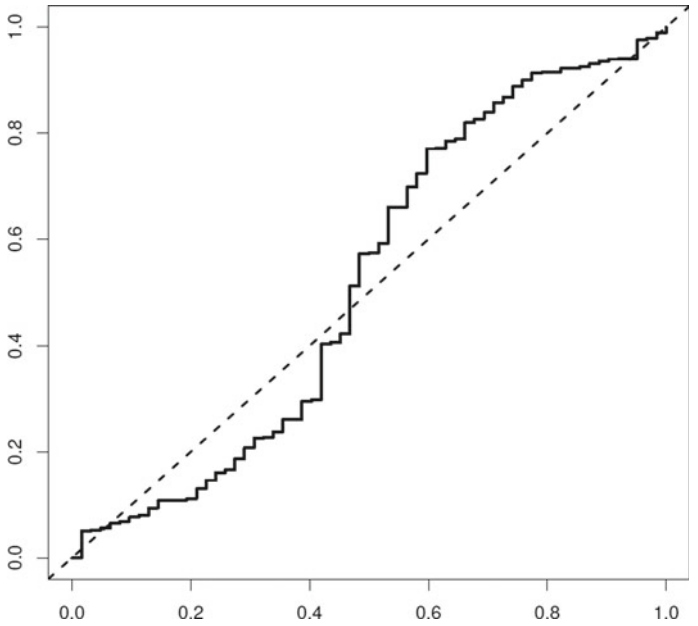


Fig. 2 Total Time on Test (TTT) plot for failure data of ROK artillery ID-1

bathtub-shaped patterns of failures are also dominant in the other artilleries in the TTT plots.

As a test for non-monotonic trends in recurrent failures, a large positive value of Vaurio's statistic [14]

$$V = \frac{\sum_{i=1}^{n} |t_i - t_n/2| - nt_n/4}{t_n\sqrt{n/48}}$$

indicates the presence of a bathtub behavior, while a large negative value indicates the presence of an inverse bathtub behavior. In applying the Vaurio's trend test to eight sets of artillery repair data, we summarized the test results in Table 1, along with their p-values. At significance level $\alpha = 0.05$, the test results provide statistical evidence of the bathtub behavior of failure intensity with respect to failure data of the eight artillery systems.

Based on the log-likelihood in (12), MLEs for S-LLP model parameters were computed using the artillery data and the details on the algorithm are described at Appendix A in Mun et al. [15]. Estimates of the S-LLP model, along with their standard errors, are given in Table 2. To obtain the standard errors of $\hat{\boldsymbol{\theta}} \equiv \left(\hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2\right)^T$, the estimated variance–covariance matrix was computed as the inverse of the estimated Fisher information matrix. For artillery ID-1, for instance, the estimated variance–covariance matrix is

Table 1 Statistical trend tests for eight sets of ROK artillery repair data

ID	Vaurio's statistic (V)	p-value
1	3.9790	< 0.0001
2	1.9863	0.0235
3	2.5615	0.0052
4	2.5967	0.0047
5	3.3462	0.0004
6	3.6172	0.0001
7	3.5558	0.0002
8	4.3812	< 0.0001

$$\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} = \begin{bmatrix} 276.0 & -12.40 & 11.80 & 26.90 \\ & 0.330 & -0.170 & -0.331 \\ & & 0.010 & 0.015 \\ & & & 0.023 \end{bmatrix} \times 10^{-6}$$

and the standard errors of $\hat{\boldsymbol{\theta}}$ are the square roots of diagonal elements in $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$. Approximate 95% confidence intervals for $\boldsymbol{\theta}$ can be constructed using the lognormal approximation.

Using the MLEs for the S-LLP model parameters, we can obtain the MLE for the expected number of failures, $\hat{\Lambda}(t)$, from (14). Figure 3 depicts $\hat{\Lambda}(t)$ under the S-PLP and the BBIP assumption, as well as $\hat{\Lambda}(t)$ under the S-LLP assumption. The figure shows that the S-LLP provides the best representation for the whole data set of eight artillery systems. Admittedly, the S-LLP model, as well as the S-PLP and the BBIP models, fails to handle the early failure data. All of the artillery repair data contain a time-lag to first failure (see Fig. 1), and it is not easy for the superposed models to represent a bathtub-shaped failure intensity that can explicitly fit the time-lag to first failure. More complex and highly parameterized models, for instance, that include the addition of a constant into the intensity functions of S-PLP, BBIP, and S-LLP, may be an alternative to capture the time-lag, but it will greatly increase model complexity as well. Under the S-LLP model, 90% (pointwise) confidence intervals for $\Lambda(t)$ are plotted for eight individual sets of artillery repair data in Fig. 4.

4 Mixed-Effects NHPP Model

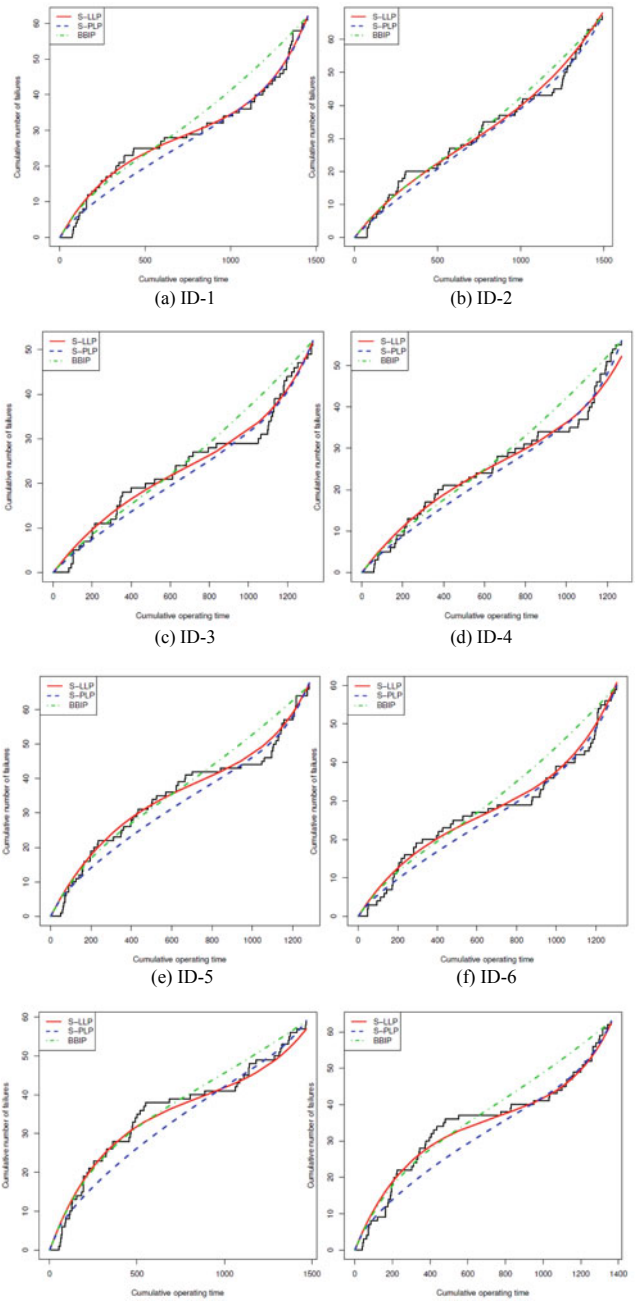
Occasionally, multiple repairable systems may present system-to-system variability due to changes in operating environments and working intensities of individual systems. In this case, it may be more reasonable to assume a heterogeneity among all the systems. Lawless [16] refers to such effects as "unobserved heterogeneity". To take the heterogeneity among systems into account, Bayesian methods (both empirical and hierarchical) have been applied to multiple repairable systems due to

Table 2 ML estimates of the S-LLP parameters and their standard errors for eight sets of artillery repair data (corresponding approximate 95% confidence intervals under the lognormal approximation in parentheses)

ID	$\hat{\alpha}_1$	$\widehat{s.e.}(\hat{\alpha}_1)$	$\hat{\beta}_1$	$\widehat{s.e.}\left(\hat{\beta}_1\right)$	$\hat{\alpha}_2$	$\widehat{s.e.}(\hat{\alpha}_2)$	$\hat{\beta}_2$	$\widehat{s.e.}\left(\hat{\beta}_2\right)$
1	0.0850	0.0166	0.0028	0.0006	0.0005	0.0001	0.0037	0.0002
	(0.0616, 0.1173)		(0.0020, 0.0039)		(0.0004, 0.0007)		(0.0035, 0.0040)	
2	0.0578	0.0147	0.0023	0.0007	0.0071	0.0012	0.0016	0.0001
	(0.0381, 0.0879)		(0.0014, 0.0039)		(0.0054, 0.0094)		(0.0014, 0.0019)	
3	0.0568	0.0117	0.0018	0.0005	0.0004	8.5×10^{-5}	0.0042	0.0002
	(0.0405, 0.0796)		(0.0012, 0.0029)		(0.0003, 0.0005)		(0.0039, 0.0045)	
4	0.0651	0.0126	0.0019	0.0005	0.0004	0.0001	0.0042	0.0002
	(0.0473, 0.0895)		(0.0012, 0.0029)		(0.0003, 0.0006)		(0.0038, 0.0046)	
5	0.1131	0.0183	0.0026	0.0004	0.0003	6.1×10^{-5}	0.0047	0.0002
	(0.0867, 0.1477)		(0.0020, 0.0034)		(0.0002, 0.0004)		(0.0044, 0.0051)	
6	0.0795	0.0164	0.0027	0.0006	0.0010	0.0002	0.0037	0.0002
	(0.0567, 0.1116)		(0.0019, 0.0040)		(0.0007, 0.0014)		(0.0034, 0.0039)	
7	0.1216	0.0200	0.0030	0.0005	0.0005	0.0001	0.0032	0.0002
	(0.0928, 0.1593)		(0.0023, 0.0040)		(0.0003, 0.0008)		(0.0029, 0.0036)	
8	0.1284	0.0214	0.0034	0.0005	0.0003	5.9e-05	0.0044	0.0002
	(0.0976, 0.1688)		(0.0026, 0.0044)		(0.0002, 0.0004)		(0.0041, 0.0047)	

their flexibility in accounting for parameter uncertainty and allowing the incorporation of a prior knowledge into the process under study (see, e.g., Hamada et al. [17], Reese et al. [18], Arab et al. [19]). System heterogeneity may be described via the prior distributions of the model parameters, however, there may also be homogeneity between individual systems. This homogeneity can be explicitly modeled by assuming common parameters in the Bayesian model. If prior distributions are unnecessarily assigned to the common parameters, the prior information employed to the common parameters can make the parameter estimation procedure more complicated. The computational complexity and the difficulty in choosing proper prior distributions have been obstacles for reliability engineers who wish to apply Bayesian methods to such practical reliability problems.

As another approach, the unobserved heterogeneity has been explicitly incorporated into the model under study in the formulation of mixed-effects model. Mixed-effects models, which is also called a "random-effects model", are widely



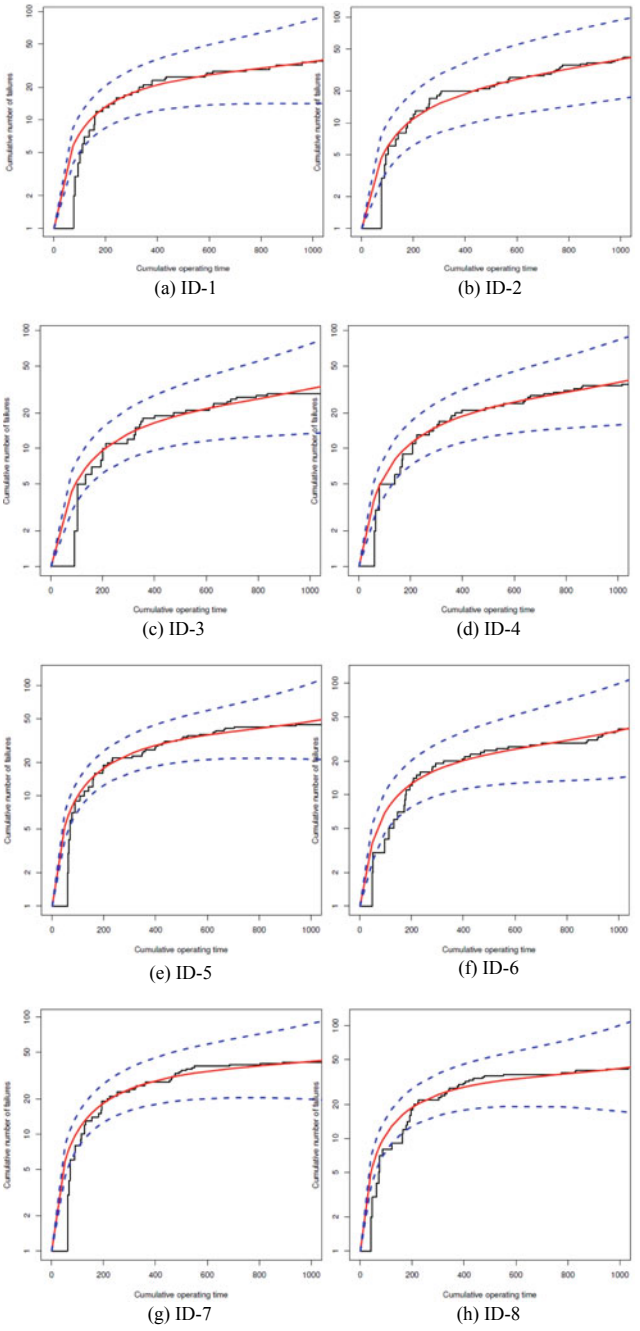
(a) ID-1 (b) ID-2

(c) ID-3 (d) ID-4

(e) ID-5 (f) ID-6

(g) ID-7 (h) ID-8

Fig. 3 Observed cumulative number of failures along with the expected number of failures $\Lambda(t)$ under the S-PLP, the BBIP, and the S-LLP assumption for eight sets of ROK artillery repair data



(a) ID-1 (b) ID-2

(c) ID-3 (d) ID-4

(e) ID-5 (f) ID-6

(g) ID-7 (h) ID-8

Fig. 4. 90% pointwise confidence intervals for $\Lambda(t)$ under the S-LLP model for eight sets of ROK artillery repair data (The vertical axis is log-scaled for better representation of the confidence intervals)

used in medical studies [20, 21], because they can model both between-individual and within-individual variation found in the data. For analyzing the reliability of multiple repairable systems, the underlying model for each individual system may be reasonably assumed to be an NHPP. Based on NHPPs with non-monotonic failure intensities, we will illustrate the inference procedure on the parameters of the mixed-effects NHPP model. The mixed-effects NHPP model allows explicit modeling and analysis of between-individual and within-individual variation of recurrent failures, along with a common baseline for all the individuals. In the formation of a mixed-effects model, the probability distributions for non-normal data involving both fixed and random effects is appropriate, a generalized mixed-effects model can be a useful tool for such purposes. The (generalized) mixed-effects models are easily implemented through commercial softwares such as S-PLUS® NLME library and SAS® NLMIXED procedure.

4.1 Mixed-Effects NHPP Model Without Covariates

Suppose that there are m independent systems; the system i is observed over the time interval $(0, T_i)$ and n_i failures are observed to occur, at times $t_{i1} < \cdots < t_{in_i}$. For the parameters $\boldsymbol{\theta}$ of the NHPP, the likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{m} \left\{ \prod_{j=1}^{n_i} \lambda\left(t_{ij}; \boldsymbol{\theta}\right) \right\} \exp\{-\Lambda(T_i; \boldsymbol{\theta})\} \tag{16}$$

with failure intensity $\lambda(\cdot)$ and its cumulative mean function $\Lambda(\cdot)$. By incorporating the inter-individual variation into the random effects $\boldsymbol{b}_i$, along with fixed effects $\boldsymbol{\zeta}$ (identical to all the systems), the conditional mean for a failure process of the ith system $\boldsymbol{t}_i = (t_{i1}, ..., t_{in})^T$ is $E[\boldsymbol{t}_i|\boldsymbol{b}_i] \equiv \boldsymbol{\mu}_i = \Lambda(\boldsymbol{t}_i|\boldsymbol{b}_i)$. The contribution to the likelihood function (16) having observed failures n_i at times t_{ij} for individual system i is

$$\mathcal{L}_i(\boldsymbol{\zeta}) = \int_{\boldsymbol{b}_i} \left\{ \prod_{j=1}^{n_i} \lambda(t_{ij}|\boldsymbol{b}_i) \right\} \exp\{-\Lambda(T_i|\boldsymbol{b}_i)\} p(\boldsymbol{b}_i) d\boldsymbol{b}_i$$

The likelihood function with parameters $\boldsymbol{\zeta}$ and $\boldsymbol{b}_i$ from the sample of m systems has the form

$$\mathcal{L}(\boldsymbol{\zeta}) = \prod_{i=1}^{m} \int_{\boldsymbol{b}_i} \left\{ \prod_{j=1}^{n_i} \lambda(t_{ij}|\boldsymbol{b}_i) \right\} \exp\{-\Lambda(T_i|\boldsymbol{b}_i)\} p(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{17}$$

and maximizing the likelihood function (17) yields the maximum likelihood estimate (MLE) of $\boldsymbol{\zeta}$, denoted by $\hat{\boldsymbol{\zeta}}$.

4.2 Mixed-Effects NHPP Model with Covariates

Suppose that individual i has a covariate vector $\boldsymbol{x}_{\mathrm{i}}$ and a failure intensity $\lambda_{\boldsymbol{x}_i}(t_{ij};\boldsymbol{\theta},\boldsymbol{\xi}_i)$, then the contribution to the likelihood function for individual i with fixed effects $\boldsymbol{\beta}_{\boldsymbol{x}}$ and random-effects $\boldsymbol{b}_i$ for $\boldsymbol{\xi}_i \equiv (\boldsymbol{\beta}_x, \boldsymbol{b}_i)^T$ is given by

$$\mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\beta}_x,\boldsymbol{b}_i) = \int_{\boldsymbol{b}_i}\left\{\prod_{j=1}^{n_i}\lambda_{\boldsymbol{x}_i}(t_{ij};\boldsymbol{\theta},(\boldsymbol{\beta}_x,\boldsymbol{b}_i))\right\}\exp\{-\Lambda_{\boldsymbol{x}_i}(T_i;\boldsymbol{\theta},(\boldsymbol{\beta}_x,\boldsymbol{b}_i))\}p(\boldsymbol{b}_i)d\boldsymbol{b}_i. \tag{18}$$

The NHPP is flexible in that the covariate information, if exists, can be explicitly modeled via the failure intensity

$$\lambda_{\boldsymbol{x}_i}(t_{ij};\boldsymbol{\theta},\boldsymbol{\xi}_i) = \lambda_0(t_{ij};\boldsymbol{\theta})h(\boldsymbol{x}_i;\boldsymbol{\xi}_i), \tag{19}$$

where $\boldsymbol{\xi}_i$ is the coefficient vector for covariate $\boldsymbol{x}_i$, and $h(\cdot)$ is a positive-valued monotonic differentiable function, e.g., $\exp(\cdot)$ or $\log(\cdot)$. The NHPP model with the failure intensity (19) is called a "proportional intensity Poisson process model" and $\lambda_0(t_{ij};\boldsymbol{\theta})$ serves as the baseline intensity function. The baseline intensity function is assumed to be constant across individuals; that is, $\boldsymbol{\theta}$ has fixed effects. Inter-individual variability is instead incorporated in the function $h(\boldsymbol{x}_i;\boldsymbol{\xi}_i)$. The model with $h(\boldsymbol{x}_i;\boldsymbol{\xi}_i) \equiv \exp(\boldsymbol{x}_i^T\boldsymbol{\xi}_i)$ has been commonly employed because it is convenient and flexible (e.g., Andersen and Gill [22]). The mean intensity function corresponding to the failure intensity (19) is $\Lambda_{\boldsymbol{x}_i}(t;\boldsymbol{\theta},\boldsymbol{\xi}_i) = \Lambda_0(t;\boldsymbol{\theta})h(\boldsymbol{x}_i;\boldsymbol{\xi}_i)$, where $\Lambda_0(t;\boldsymbol{\theta}) = \int_0^t \lambda_0(u;\boldsymbol{\theta})\,\mathrm{d}u$. The likelihood function (18) can be rewritten by the factorization as (Cox and Lewis [5], Sect. 5.3)

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\beta}_x,\boldsymbol{b}_i) = \prod_{j=1}^{n_i}\left\{\frac{\lambda_0(t_{ij};\boldsymbol{\theta})}{\Lambda_0(t_{ij};\boldsymbol{\theta})}\right\} \times \int_{\boldsymbol{b}_i}\left\{\Lambda_0(T_i;\boldsymbol{\theta})h(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_x+\boldsymbol{b}_{\mathrm{i}}))\right\}^{n_i}\\ \exp\{-\Lambda_0(T_i;\boldsymbol{\theta})h(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_x+\boldsymbol{b}_{\mathrm{i}}))\}p(\boldsymbol{b}_i)d\boldsymbol{b}_i.\end{aligned} \tag{20}$$

The likelihood function for a sample of m independent individuals is the product of terms $\mathcal{L}_1,\ldots,\mathcal{L}_m$ giving

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta}_x) = \mathcal{L}_1(\boldsymbol{\theta})\mathcal{L}_2(\boldsymbol{\theta},\boldsymbol{\beta}_x), \tag{21}$$

where $\mathcal{L}_1(\boldsymbol{\theta})$ is the product of the first terms and $\mathcal{L}_2(\boldsymbol{\theta},\boldsymbol{\beta}_x)$ is the product of the second terms in right-hand side of (20). Lawless [16] considered the following intensity function

$$\lambda_{\boldsymbol{x}}(t_{ij};\theta,\boldsymbol{\beta}_x,\boldsymbol{b}_i) = \lambda_0(t_{ij};\boldsymbol{\theta})b_i exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_x), \tag{22}$$

where $\boldsymbol{b}_i$ is assumed to be an iid gamma-distributed random variable with mean 1 and variance ϕ. In this case the second term in (20) becomes

$$\frac{\Gamma\left(n_i+\phi^{-1}\right)}{\Gamma\left(\phi^{-1}\right)} \cdot \frac{\left[\phi\Lambda_0(\tau_i;\boldsymbol{\theta})\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_{\boldsymbol{x}}\right)\right]^{n_i}}{\left[1+\phi\Lambda_0(\tau_i;\boldsymbol{\theta})\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_{\boldsymbol{x}}\right)\right]^{n_i+\phi^{-1}}},$$

which is a negative binomial regression model. The negative binomial model is a reasonable model to accommodate extra-Poisson variability. For instance, if the baseline intensity function has a power law process, for $\boldsymbol{\theta} \equiv (\alpha, \beta)$

$$\mathcal{L}_1(\boldsymbol{\theta}) = \prod_{i=1}^{m}\prod_{j=1}^{n_i}\left(\frac{\beta}{t_{ij}}\right) \quad and \quad \mathcal{L}_2\left(\boldsymbol{\theta},\boldsymbol{\beta}_{\boldsymbol{x}}\right) = \prod_{i=1}^{m}\frac{\Gamma\left(n_i+\phi^{-1}\right)}{\Gamma\left(\phi^{-1}\right)} \cdot \frac{\left[\phi(T_i/\alpha)^{\beta}\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_{\boldsymbol{x}}\right)\right]^{n_i}}{\left[1+\phi(T_i/\alpha)^{\beta}\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_{\boldsymbol{x}}\right)\right]^{n_i+\phi^{-1}}}, \tag{23}$$

4.3 Estimation of Parameters in Mixed-Effects NHPP Model

In general, the integral calculations in the likelihood function (17) and (21) involve high-dimensional integration, and do not produce closed-form expressions, requiring numerical integration techniques to estimate the likelihood function. Bae and Kvam [23] introduced various approximation methods to numerically optimize the likelihood function from repeated-measured degradation data of vacuum fluorescent displays when the distribution of $\boldsymbol{b}_i$ is multivariate normal. SAS® NLMIXED procedure provides several approximation methods including adaptive Gaussian quadrature [24] and first-order method [25] for the mixed-effects model.

In the NHPP model without covariates, ML estimates of $\boldsymbol{\zeta}$ are obtained by maximizing the likelihood function (17) numerically or using approximation methods (if necessary). A simple approach to estimation in NHPP model with covariates is to estimate $\boldsymbol{\theta}$ by maximizing $\mathcal{L}_1(\boldsymbol{\theta})$, and then to maximize (21) with respect to $\boldsymbol{\beta}_{\boldsymbol{x}}$, with $\boldsymbol{\theta}$ fixed at their estimates [14]. With the PLP baseline intensity, for example, we can first estimate β in the likelihood function (23) by maximizing $\mathcal{L}_1$ with respect to β, then plug in $\hat{\beta}$ and maximize $\mathcal{L}_2$ with respect to ϕ, α, and $\boldsymbol{\beta}_{\boldsymbol{x}}$. Maximization of $\mathcal{L}_2$ for fixed β is easy using Newton's method or the scoring algorithm [26].

The random-effects in the mixed-effects NHPP model are assumed to have normal distributions with zero means. Their specific values for a given individual are just realizations from the normal distributions. These random effects can be efficiently estimated using empirical Bayes methods [27]. For the failure process of the ith system $\boldsymbol{t}_i$, empirical Bayes estimates of $\boldsymbol{b}_i$ (denoted by $\hat{\boldsymbol{b}}_i$) is given by the posterior mean of $\boldsymbol{b}_i$ as

$$\hat{\boldsymbol{b}}_i = E(\boldsymbol{b}_i|\boldsymbol{t}_i) = \frac{\int_{\boldsymbol{b}_i}\boldsymbol{b}_i\, p(\boldsymbol{t}_i|\boldsymbol{b}_i)p(\boldsymbol{b}_i)d\boldsymbol{b}_i}{\int_{\boldsymbol{b}_i} p(\boldsymbol{t}_i|\boldsymbol{b}_i)p(\boldsymbol{b}_i)d\boldsymbol{b}_i}$$

for the conditional probability function of $\boldsymbol{t}_i$ given $\boldsymbol{b}_i$, $p(\boldsymbol{t}_i|\boldsymbol{b}_i)$. If parametric assumptions on the distribution of random-effects are made, e.g., normal, then empirical Bayes methods are equivalent to best linear unbiased prediction (BLUP) methods [28].

Confidence intervals can be constructed for the parameters of the mixed-effects model or their functions based on standard errors derived from the (observed) Fisher information matrix. In generalized mixed-effects NHPP model without covariates, a large-sample approximation of standard errors of the ML estimators is given through the estimated variance–covariance matrix $\hat{\Xi}_{\hat{\zeta}}$, which is computed as the inverse of the observed Fisher information matrix. That is, $\hat{\Xi}_{\hat{\zeta}} \equiv \mathcal{I}\left(\hat{\zeta}\right)^{-1}$ for $\mathcal{I}\left(\hat{\boldsymbol{\zeta}}\right) = -\partial^2 l/\partial\boldsymbol{\zeta}^2$ evaluated at $\boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}}$, where $l = \log \mathcal{L}(\boldsymbol{\zeta})$. For example, in the NHPP model with covariates, the asymptotic variance–covariance matrix of $\left(\boldsymbol{\theta}, \boldsymbol{\beta}_{\boldsymbol{x}}\right)$ is obtained as $\mathcal{I}\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{x}}\right)^{-1}$, where

$$\mathcal{I}\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{x}}\right)^{-1} \equiv \begin{bmatrix} -\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2} & -\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\beta}_{\boldsymbol{x}}} \\ & -\frac{\partial^2 l}{\partial \boldsymbol{\beta}_{\boldsymbol{x}}^2} \end{bmatrix}^{-1} = \begin{bmatrix} -\left(\frac{\partial^2 l_1}{\partial \boldsymbol{\theta}^2} + \frac{\partial^2 l_2}{\partial \boldsymbol{\theta}^2}\right) & -\frac{\partial^2 l_2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\beta}_{\boldsymbol{x}}} \\ & -\frac{\partial^2 l_2}{\partial \boldsymbol{\beta}_{\boldsymbol{x}}^2} \end{bmatrix}^{-1}$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\beta}_{\boldsymbol{x}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{x}}$. Here, $l_1 = \log \mathcal{L}_1(\boldsymbol{\theta})$ and $l_2 = \log \mathcal{L}_2\left(\boldsymbol{\theta}, \boldsymbol{\beta}_{\boldsymbol{x}}\right)$. Then, similar to the case of the NHPP model without covariates, approximate standard errors for (or functions of) $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}}$ are computed using the delta method, and their Wald-type confidence intervals are also computed. For the random-effects, the standard errors of $\hat{\boldsymbol{b}}_i$ are computed using the delta method and confidence intervals of the random-effects may be constructed using the Wald-type statistics.

After fitting mixed-effects NHPP model to failure-time data from multiple repairable systems, we need to assess the significance of the terms in the model. The significance test can be done through a likelihood ratio statistic. Denote $\mathcal{L}_F$ as the likelihood for the full model, and $\mathcal{L}_R$ as the likelihood for the reduced model. Then under the null hypothesis that the reduced model is adequate, the likelihood ratio test (LRT) statistic

$$2\log(\mathcal{L}_F/\mathcal{L}_R) = 2(\log \mathcal{L}_F - \log \mathcal{L}_R)$$

will approximately follow a χ^2 distribution with $(\psi_F - \psi_R)$ degrees of freedom, where ψ_F and ψ_R are the number of parameters to be estimated in the full and reduced model, respectively.

Even though the LRT can assess the significance of particular terms, model selection procedure via such pairwise comparisons has been criticized owing to an overuse of hypothesis testing. By contrast, an information-based model selection procedure allows comparison of multiple candidate models. Two widely used information criteria for assessing model fit are Akaike's information criterion (AIC) [29] and the Bayesian information criterion (BIC) [30]. For the log-likelihood of a model, l,

the AIC and BIC are, respectively

$$\text{AIC} = -2l + 2p^*, \quad \text{and} \quad \text{BIC} = -2l + p^* \log N$$

where p^* denotes the total number of parameters in the model, and N denotes the total number of observations in the data set; that is, $N = \sum_{i=1}^{\cdot} n_i$ for the mixed-effects NHPP model. If we use the AIC to compare several models for the same data, we prefer the model with the lowest AIC value. Similarly, when using the BIC, we prefer the model with the lowest BIC value.

Residuals can be set up to provide checks on the assumed model. Under the NHPP model, the quantities $\Lambda(t_{ij}) - \Lambda(t_{i,j-1})$ are independent standard exponential random variables for $j = 1, ..., n_i$. Therefore, residuals $e_{ij} = \hat{\Lambda}(t_{ij}) - \hat{\Lambda}(t_{i,j-1})$ should look like standard exponential random variables if the NHPP model under assumptions is correct. The deviation from the model assumptions can be checked by plotting $(e_{ij}, e_{i,j-1})$ to detect serial correlation with respect to j in the e_{ij}'s. See Lawless [14] for more details on the properties of residuals and formal model assessment using the residuals.

4.4 *Application of Mixed-Effects NHPP Model to Artillery Repair Data*

Mun et al. [15] analyzed field-repair data of eight sets of artillery systems where their failure intensities appear bathtub-shaped. Failure frequency also tended to vary greatly across all the systems. Mun et al. [15] proposed the S-LLP model (7) instead of S-PLP model proposed by Pulcini [9] to describe the artillery repair data with bathtub-shaped failure intensity. To incorporate individual variability into the superposed NHPP models, we considered both the mixed-effects S-PLP model and the mixed-effects S-LLP model. For the mixed-effects S-PLP model, the general model for comparison is a mean failure intensity of the S-PLP with four random-effects

$$\Lambda_{ij}(t) = \left(\frac{t_{ij}}{\zeta_1 + b_{i1}}\right)^{\vartheta_1 + b_{i2}} + \left(\frac{t_{ij}}{\zeta_2 + b_{i3}}\right)^{\vartheta_2 + b_{i4}},$$

and similarly, the general model for the mixed-effects S-LLP is

$$\Lambda_{ij}(t) = \left(\frac{\gamma_1 + b_{i1}}{\kappa_1 + b_{i2}}\right)\left(1 - e^{-(\kappa_1 + b_{i2})t_{ij}}\right) + \left(\frac{\gamma_2 + b_{i3}}{\kappa_2 + b_{i4}}\right)\left(e^{(\kappa_2 + b_{i4})t_{ij}} - 1\right),$$

where the random-effects $(b_{i1}, b_{i2}, b_{i3}, b_{i4})$ of the two models have general covariance structures, for $i = 1, ..., m$, $j = 1, ..., n_i$. After executing

the LRT procedure and computing the AIC and BIC, the final parameter estimates of mixed-effects S-PLP model are: $\hat{\zeta}_1 = 4.1720, \hat{\vartheta}_1 = 0.6544, \hat{\zeta}_2 = 1095.51, \hat{\vartheta}_2 = 12.5636.$, and $(b_{i1}, b_{i2}, b_{i3}, b_{i4})^T \sim \mathcal{N}(0, 0, 0, 0)^T$, diag(4.8261, 0.0028, 1.6880 × 104, 14.0611), where diag(·) denotes a diagonal matrix. The final parameter estimates of the mixed-effects S-LLP model are: $\hat{\gamma}_1 = 0.0843, \hat{\kappa}_1 = 0.0020, \hat{\gamma}_2 = 5.4120 \times 10^{-5}, \hat{\kappa}_2 = 0.0057$, and $(b_{i1}, b_{i2})^T \sim \mathcal{N}(0, 0)^{\mathrm{T}}$, $\mathrm{diag}\left(4.8018 \times 10^{-4}, 2.0789 \times 10^{-7}\right)$, and $b_{i4} \sim \mathcal{N}\left(0, 2.0737 \times 10^{-7}\right)$.

We compared their modeling performance with individually fitted S-PLP and S-LLP models, correspondingly, in terms of mean square errors. Before comparing their modeling performance, we performed the diagnostics for the fitted models based on the residuals derived from each of the super- posed NHPP models. The histograms of the residuals from the fitted models (Fig. 5) justify the assumptions for the four superposed NHPP models. Each of superposed NHPP models incorporating both fixed-effects and random-effects has smaller MSE than individually fitted superposed NHPP models (see Table 3). We chose the mixed-effects S-LLP model which has the smallest average MSE with respect to artillery systems data for further analytical purpose. Table 4 compares parameter estimates of individually fitted S-LLP model and those of mixed-effects S-LLP model and their 95% pointwise confidence intervals using the lognormal approximation for the eight artillery systems. The parameter estimates of mixed-effects S-LLP model are consistently smaller than those of individually fitted S-LLP model, and their confidence intervals are consistently shorter than those of individually fitted S-LLP model. We also observed that mixed-effects S-PLP model has consistently shorter confidence intervals than individually fitted S-PLP model. The estimate of cumulative number of failures and its 95% (pointwise) confidence intervals are plotted for eight individual sets of artillery systems data in Fig. 6.

5 Conclusions

Some complex systems show the bathtub-shaped failure intensity that are characterized by a number of different failure modes as in the repairable artillery systems. The monotonic failure intensity models such as the PLP and the LLP models are not appropriate to model the bathtub-shaped failure pattern. As an alternative, a superposed log linear process (S-LLP), which is a mixture of nonhomogeneous Poisson processes, was developed to model this kind of non-monotonic failure intensity. The derived S-LLP model is shown to be much better at fitting the repair data than previous models that have been derived for bathtub-shaped failure intensities. Although the estimation problem is computationally cumbersome, the MLEs are straightforward and can be used to construct approximate confidence bounds for cumulative failure intensity.

For multiple repairable systems presenting system-to-system variability owing to operation environments or working intensities of individual systems, we go over the

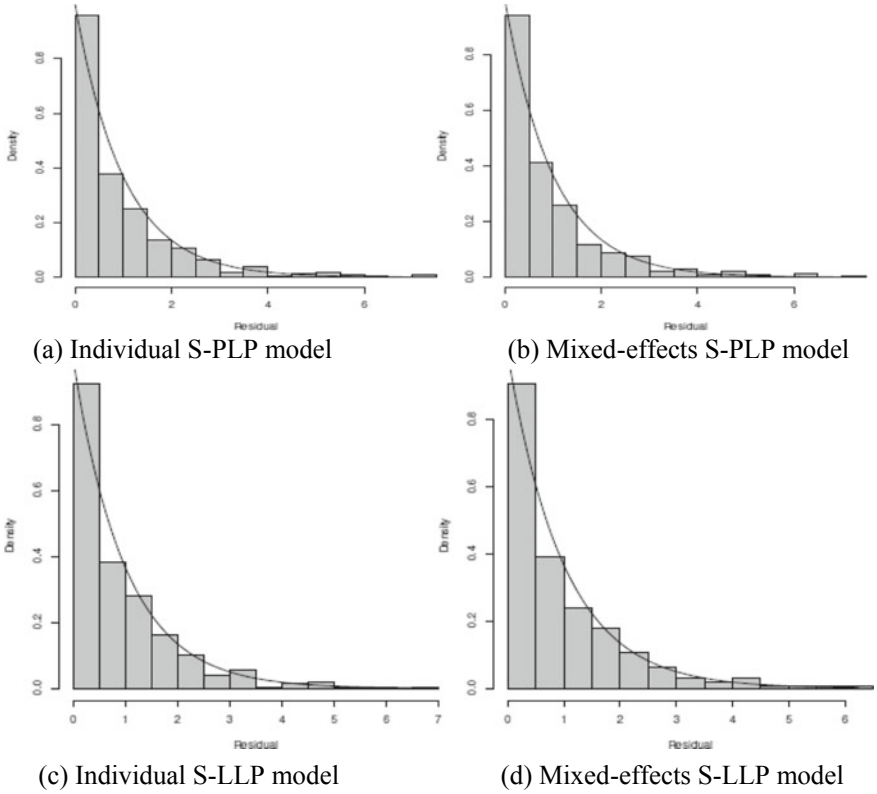


Fig. 5 Histograms of the residuals from each of the superposed NHPP models for eight artillery systems

Table 3 Mean squared errors between observed and estimated number of failures from each of superposed NHPP models for eight artillery systems

ID	S-PLP	S-PLP-NLMM	S-LLP	S-LLP-NLMM
ID-1	9.8153	5.0171	3.1279	2.7201
ID-2	6.4003	3.1235	3.5983	2.7279
ID-3	7.3453	4.2244	3.4652	3.2335
ID-4	6.5224	3.2255	3.7683	2.1790
ID-5	14.2063	6.8818	4.7604	3.4388
ID-6	8.2261	4.4996	3.4576	3.1208
ID-7	19.4709	9.6250	5.1072	2.9708
ID-8	21.1419	10.6802	6.1606	4.3777
Total	11.7504	5.9515	4.2018	3.1076

Table 4 Parameter estimates of both individually fitted S-LLP model and mixed-effects S-LLP model, along with their approximate 95% confidence intervals under the lognormal approximation in parentheses

	S-LLP				S-LLP-NLMM			
ID	$\hat{\gamma}_1$	$\hat{\kappa}_1$	$\hat{\gamma}_2(\times 10^{-3})$	$\hat{\kappa}_2(\times 10^{-3})$	$\hat{\gamma}_1$	$\hat{\kappa}_1(\times 10^{-3})$	$\hat{\gamma}_2(\times 10^{-3})$	$\hat{\kappa}_2(\times 10^{-3})$
ID-1	0.085	0.003	0.538	3.725	0.079	2.108	5.412	5.521
	(0.045, 0.160)	(0.001, 0.007)	(0.011, 27.337)	(1.653, 8.390)	(0.046, 0.136)	(1.380, 3.221)	(5.409, 5.415)	(4.697, 6.490)
ID-2	0.060	0.002	5.481	1.728	0.064	1.152	5.412	5.255
	(0.028, 0.128)	(0.001, 0.008)	(0.836, 35.944)	(0.770, 3.881)	(0.032, 0.125)	(0.530, 2.502)	(5.409, 5.415)	(4.434, 6.228)
ID-3	0.057	0.002	0.370	4.205	0.057	1.508	5.412	5.777
	(0.029, 0.112)	(0.001, 0.006)	(0.007, 20.002)	(1.918, 9.218)	(0.027, 0.121)	(0.834, 2.727)	(5.409, 5.415)	(4.950, 6.742)
ID-4	0.063	0.002	0.581	4.085	0.067	1.754	5.412	6.264
	(0.033, 0.120)	(0.001, 0.005)	(0.034, 9.845)	(2.243, 7.440)	(0.035, 0.127)	(1.053, 2.919)	(5.409, 5.415)	(5.432, 7.223)
ID-5	0.114	0.003	0.824	3.742	0.108	2.248	5.412	6.119
	(0.068, 0.191)	(0.002, 0.005)	(0.042, 16.176)	(1.854, 7.551)	(0.073, 0.161)	(1.511, 3.346)	(5.409, 5.415)	(5.288, 7.080)
ID-6	0.079	0.003	0.635	4.042	0.075	1.943	5.412	6.151
	(0.042, 0.152)	(0.001, 0.006)	(0.021, 19.151)	(1.966, 8.308)	(0.042, 0.133)	(1.227, 3.078)	(5.409, 5.415)	(5.320, 7.111)
ID-7	0.122	0.003	0.518	3.196	0.111	2.351	5.412	4.865
	(0.074, 0.201)	(0.002, 0.005)	(0.084, 3.188)	(2.032, 5.028)	(0.075, 0.163)	(1.607, 3.438)	(5.409, 5.415)	(4.050, 5.845)
ID-8	0.133	0.003	0.437	3.930	0.114	2.571	5.412	5.651
	(0.080, 0.223)	(0.002, 0.007)	(0.083, 2.308)	(2.732, 5.655)	(0.079, 0.166)	(1.816, 3.639)	(5.409, 5.415)	(4.825, 6.618)

application of mixed-effects models to recurrent failure data from multiple repairable systems based on the superposed Poisson process to model the bathtub-shaped failure intensities. The mixed-effects models explicitly involve between-system variation through random-effects, along with a common baseline for all the systems through fixed-effects for both normal and non-normal data. Details on estimation of the parameters of the mixed-effects superposed Poisson process models and construction of their confidence intervals are examined. An applicative example shows prominent proof of the mixed-effects superposed Poisson process models for the purpose of reliability analysis.

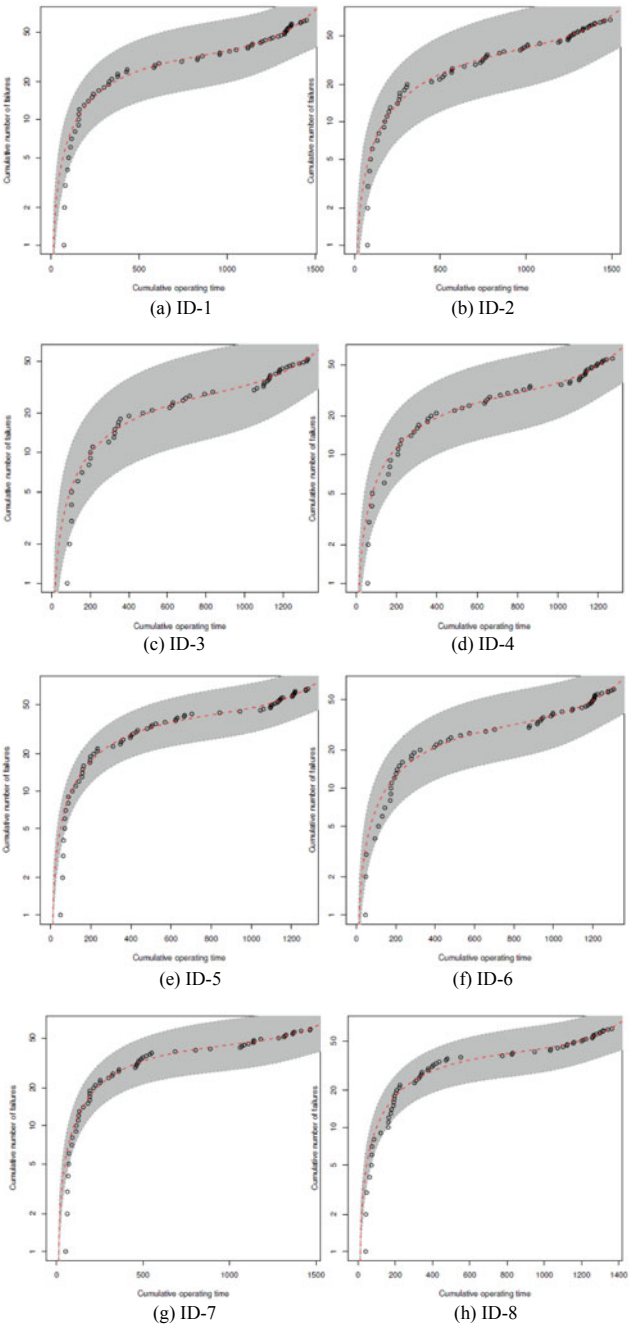


Fig. 6 $\hat{\Lambda}(t)$ and 95% pointwise confidence intervals for $\hat{\Lambda}(t)$ under the mixed-effects S-LLP model for eight sets of artillery systems (The vertical axis is log-scaled for better representation of the confidence intervals)

References

1. Ascher H, Feingold H (1984) Repairable systems reliability: modeling, inference, misconceptions and their causes. Marcel Dekker, New York
2. Rigdon SE, Basu AP (2000) Statistical methods for the reliability of repairable systems. Wiley, New York
3. Krivtsov VV (2007) Practical extensions to NHPP application in repairable system reliability analysis. Reliab Eng Syst Saf 92(5):560–562
4. Crow LH (1975) Reliability analysis for complex, repairable systems. Army materiel systems analysis activity aberdeen proving ground Md
5. Cox DR, Lewis PA (1966) Statistical analysis of series of events. Methuen, London
6. Pulcini G (2001) A bounded intensity process for the reliability of repairable equipment. J Qual Technol 33(4):480–492
7. Nelson W (1988) Graphical analysis of system repair data. J Qual Technol 20(1):24–35
8. Musa JD, Okumoto K (1984) A logarithmic poisson execution time model for software reliability measurement. In: Proceedings of the 7th international conference on software engineering, pp 230–238
9. Pulcini G (2001) Modeling the failure data of a repairable equipment with bathtub type failure intensity. Reliab Eng Syst Saf 71(2):209–218
10. Yang TY, Kuo L (1999) Bayesian computation for the superposition of nonhomogeneous poisson processes. Can J Stat 27(3):547–556
11. Hjorth U (1980) A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rates. Technometrics 22(1):99–107
12. Guida M, Pulcini G (2009) Reliability analysis of mechanical systems with bounded and bathtub shaped intensity function. IEEE Trans Reliab 58(3):432–443
13. Barlow RE, Davis B (1977) Analysis of time between failures for repairable components. In: Nuclear systems reliability engineering and risk assessment. SIAM, Philadelphia, pp 543–561
14. Vaurio JK (1999) Identification of process and distribution characteristics by testing monotonic and non-monotonic trends in failure intensities and Hazard rates. Reliab Eng Syst Saf 64:345–357
15. Mun BM, Bae SJ, Kvam PH (2013) A superposed log-linear failure intensity model for repairable artillery systems. J Qual Technol 45:100–115
16. Lawless JF (1987) Regression methods for poisson process data. J Am Stat Assoc 82:808–815
17. Hamada MS, Wilson AG, Reese CS, Martz HF (2008) Bayesian reliability. Wiley, New York
18. Reese CS, Wilson AG, Guo J, Hamada MS, Johnson VE (2011) A Bayesian model for integrating multiple sources of lifetime information in system reliability assessement. J Qual Technol 43:127–141
19. Arab A, Rigdon SE, Basu AP (2012) Bayesian inference for the piecewise exponential model for the reliability of multiple repairable systems. J Qual Technol 44:28–38
20. Verbeke G, Lesaffre E (1996) A linear mixed-effects model with heterogeneity in the random-effects population. J Am Stat Assoc 91:217–221
21. Hartford A, Davidian M (2000) Consequneces of misspecifying assumptions in nonlinear mixed effects model. Comput Stat Data Anal 34:139–164
22. Anderson PK, Gill R (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10:1100–1120
23. Bae SJ, Kvam PH (2004) A nonlinear random coefficients model for degradation testing. Technometrics 46:460–469
24. Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat 4:12–35
25. Beal SL, Sheiner LB (1982) Estimating population kinetics. CRC Crit Rev Biomed Eng 8:195–222
26. Mccullagh P, Nelder JA (1989) Generalized linear model. Chapman and Hall, New York
27. Everitt B, Howell D (2005) Encyclopedia of statistics in behavioral science. Wiley, New York

28. Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Stat Sci 6:15–32
29. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723
30. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Multi-state Signatures for Multi-state Systems with Binary/Multi-state Components

He Yi and Narayanaswamy Balakrishnan

Abstract Signature theory, as an important part of reliability theory, provides an efficient tool for modeling and analyzing various properties of reliability systems. By now, signature theory has become exhaustive for binary-state systems, but for multi-state systems which are commonly encountered in practice, there are still a lot of issues to examine. In this work, we review important research works that have been carried out on signatures of systems with a special focus on signature concepts, their properties, computational methods, and some multi-state signatures for multi-state systems. We also summarize work that we have done recently on multi-state signatures, including their definitions, properties, transformation formulas and module structures. Finally, we present a number of examples to illustrate various notations and associated results described here.

Keywords Signature · Reliability · Multi-state systems · Multi-state signatures · Modeling and analysis

1 Motivation

Signature theory plays a very important role in the field of reliability since it was first introduced by Samaniego [59]. Reliability function of a coherent system can be represented through its system signature and its common component lifetime distribution; this is useful not only in the reliability calculation of large practical systems, but also for developing statistical inference for component as well as systemlifetime distributions. Next, orderings of signatures (such as stochastic ordering,

H. Yi (✉)
School of Economics and Management, Beijing University of Chemical Technology, North Third Ring Road 15, Chaoyang District 100029, Beijing, China
e-mail: yihe@mail.buct.edu.cn

N. Balakrishnan
Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada
e-mail: bala@mcmaster.ca

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_8

hazard rate ordering and likelihood ordering) result in corresponding orderings of system lifetimes, which facilitate lifetime comparisons of reliability systems of the same or different sizes with the use of transformation formulas between signatures of different sizes. In addition, when a reliability system is working at a certain time, signature of the system can also be used to study residual lifetime and stochastic comparisons of the system, regardless of whether the number of failed components is known or not.

For reliability systems, there are a lot of measures to describe their structural properties, of which the structure function is the most common one. However, for some complex systems, especially large network systems, their structure functions tend to become very complex, which renders further research and applications to be quite difficult. In contrast to the structure function, system signature is a vector with the same dimension as the number of components in the system, and it possesses clear probabilistic interpretation and is completely determined by the system structure. System signature can easily be applied and computed for practical systems, which makes it to be one of the most important measures to describe system structure and is therefore widely used in the fields of reliability and network. Currently, the signature theory for binary-state systems seems to be quite advanced, including definitions of different signature concepts, properties and applications of these concepts, computational methods for different signature concepts, statistical inference for component as well as system lifetime distributions, and so on.

Multi-state systems have become a subject of great interest recently. Due to factors such as aging, a reliability system in practice often goes through some imperfect functioning states before entering a complete failure state from the initial perfect functioning state, and moreover may often have more than one failure mode. For these reasons, multi-state systems exist commonly in practice, and traditional theories and methods for binary-state systems can not meet the requirements for the reliability analysis of such practical multi-state systems. Naturally, it will be quite useful if signature theory can be generalized from binary-state systems to multi-state systems, and it seems that there is still a lot of work to be done in this regard. For example, more signature concepts need to be defined for multi-state systems, especially those that depend only on specific system structures. Then, their stochastic properties, computational methods and transformation formulas can all be subsequently developed to analyze the corresponding multi-state reliability systems.

2 Introduction

The concept of system signature was first put forward by Samaniego [59] in 1985, and then formally named as such by Kochar et al. [29]. It can be readily seen from the published literature that signatures have received considerable attention in the fields of reliability and network, as well as the associated theoretical research on order statistics and distribution theory [45]. There are many influential monographs

in the field of signature, such as [6, 24, 34, 60]. In this paper, we provide a review of signature theory with respect to four main topics—signature related concepts, properties of signatures, computational methods for signatures and multi-state signatures for multi-state systems. This review also emphasises on some recent developments and their possible applications to the analysis of some practical reliability systems.

2.1 Signature Related Concepts

System signature, proposed originally by Samaniego [59], can be defined in two ways: from the view of component failure orderings, or from probability viewpoint. The former one is called structural signature, which focuses on the relationship of signature and system structure, while the latter is called probabilistic signature, which yields reliability function representation of a system. For binary-state coherent systems with independent and identically distributed (*i.i.d.*) components, the two definitions are equivalent. Based on this equivalency property, Elperin et al. [20] proposed a Monte Carlo algorithm for calculating the system reliability of network systems, wherein system signature is used and referred to as internal distribution. Since then, numerous signature concepts have been proposed (see Table 1 for a list), most of which are distribution-free and depend only on system structures, except some conditional signatures [36].

For systems with exchangeable components, Navarro et al. [50] proposed minimal signature and maximal signature, based on which the upper and lower bounds of system reliability were obtained. For network systems, Gertsbakh and Shpungin [24] defined D-spectrum $(f_1, \ldots, f_n)$ and C*-spectrum $(x_1, \ldots, x_n)$ based on components failed and components functioning, respectively, and these authors also defined the corresponding cumulative D-spectrum $(f_1, f_1 + f_2, \ldots, \sum_{i=1}^{n} f_i)$, cumulative C*-spectrum $(x_1, x_1 + x_2, \ldots, \sum_{i=1}^{n} x_i)$ and BIM spectrum.

When the number of failed components is known and the system is still functioning, for studying the reliability representation and stochastic comparison of remaining lifetimes of systems, Samaniego et al. [61] introduced the concept of dynamic signature. In this regard, under more general conditions, Mahmoudi and Asadi [36] proposed conditional signature. For two coherent systems with shared components, Navarro et al. [52] introduced the notion of joint signature for the first time, and Zarezadeh et al. [76] extended it from two systems to a more general case and called it generalized joint signature. Mohammadi [44] similarly defined the joint reliability signature of two or more systems. However, these concepts have no unified expression and need to be discussed case by case. Moreover, the elements involved in these concepts have no probabilistic meaning. Therefore, Navarro et al. [54] redefined the joint signature to give it a unified form and a probabilistic meaning.

Based on the signature $\boldsymbol{s} = (s_1, \ldots, s_n)$, Gertsbakh et al. [26] defined the concepts of cumulative signature $\boldsymbol{S} = (s_1, s_1 + s_2, \ldots, \sum_{i=1}^{n} s_i)$ and tail signature $\bar{\boldsymbol{S}} = (\sum_{i=2}^{n} s_i, \ldots, \sum_{i=n}^{n} s_i, 0)$. Coolen and Coolen-Maturi [11] further extended the tail signature to systems with multiple types of components, and presented the definition

Table 1 Definitions of signature concepts

Field	Concept	Definition
Reliability	Structural signature	A vector of proportions of component failure orderings such that the ith component failure causes the failure of the system [59].
	Probabilistic signature	A vector of probabilities such that the ith component failure causes the failure of the system [59]
	Minimal/maximal signature	A vector of coefficients when the reliability function is represented as a general mixture of reliabilities of series/parallel systems of different sizes [50]
	Dynamic signature	A vector of conditional probabilities such that the ith component failure causes the failure of the system, given the number of failed components [61]
	Conditional signature	A vector of conditional probabilities such that the ith component failure causes the failure of the system, given the interval of system lifetime [36]
	Generalized joint signature	A matrix of coefficients if the joint lifetime distribution of systems is given as a mixture of products of ordered component lifetime distributions [76]
	Joint signature	A matrix of probabilities such that the ith and jth component failures cause systems 1 and 2 to fail, respectively [54]
	Joint reliability signature	A matrix of coefficients if the joint reliability of several systems is given as a mixture of products of reliabilities of ordered components [44]
	Cumulative signature	A vector of probabilities such that the system is failed at the time of the ith component failure [26]
	Tail signature	A vector of probabilities such that the system is working at the time of the ith component failure [26]
	Survival signature	A matrix of conditional probabilities such that the system is functioning, given the numbers of different types of working components [11]
	Ordered signature	The signature of the qth failed system in a life test of several independent and identical/non-identical coherent systems [5, 68]
	M-signature	A vector of probabilities such that the kth component failure in component subset M causes the failure of the system [37]
	Subsignature	M-signature of any component subset M [37]
	Joint survival signature	A collection of probabilities such that both systems function at given times, given the numbers of working components in each type [12]; see also [4]
Network	Internal distribution	The same as structural signature [20]
	D-spectrum	The same as structural signature and internal distribution [24]
	C*-spectrum	A vector of proportions of edge working orderings such that the ith edge working causes the network to be in UP state [24]
	Cumulative D/C*-spectrum	Cumulative spectrum for D-spectrum/C*-spectrum [24]
	BIM spectrum	A collection of numbers of permutations such that the first i edges (including edge e_j) construct an UP state of the network [24]

of survival signature. For the purpose of saving time and cost in a life test, it is often not necessary to keep testing until all systems fail and from this viewpoint, Balakrishnan and Volterman [5] proposed the notion of ordered signature for independent and identical coherent systems in a life test. For semi-coherent systems, Marichal [37] put forward M-signature and sub-signature, where M is a non-empty subset of the component set and it often does not contain highly reliable components.

2.2 Properties of Signatures

The concept of signature was originally proposed for systems with *i.i.d.* components, but with several subsequent developments, this assumption has been gradually weakened, with significant focus on the case of non-*i.i.d.* components.

Systems with *i.i.d.*components: Samaniego [59] gave a reliability function representation based on system signature, and some discussions on closure properties of lifetime distribution classes were also presented. Navarro et al. [46] studied the reliability function representation based on signature for system remaining lifetimes and stochastic comparisons between them. Kochar et al. [29] presented sufficient conditions on signature for system lifetimes to have stochastic ordering, hazard rate ordering and likelihood ordering. Block et al. [7] showed that these sufficient conditions are not necessary, and gave the necessary and sufficient conditions. Based on signature, Boland [8] also discussed the symmetry of indirect voting systems and compared their lifetimes with direct voting systems. Boland and Samaniego [9] proved that for consecutive-2-out-of-n: F systems, their lifetimes decrease in a stochstic ordering with an increase in the component number n. Navarro et al. [55] derived transformation formulas of signatures for systems of different sizes through the triangle rule of order statistics, while Lindqvist et al. [33] discussed the equivalency of systems of different sizes by using cut sets. Mohammadi [42, 43] classified the generalized joint signatures of m parallel systems and discussed the transformation between different classes. Amini-Seresht et al. [1] found sufficient conditions for system lifetime to be larger with components replaced by better ones in the sense of hazard rate, reverse hazard rate and likelihood orderings.

Systems with exchangeable components: Navarro et al. [49] and Navarro et al. [55] derived sufficient conditions on signature for system lifetimes to have stochastic, hazard rate and likelihood orderings. Koutras et al. [30] derived necessary and sufficient conditions for the system to have hazard rate and inverse hazard rate orderings. Navarro and Rychlik [51] constructed a maximal convex/minimal concave function by scaling the system signature, giving a reliability function representation and upper/lower bounds of the expected system lifetime. Marichal et al. [41] proved that the exchangeability of components is a necessary and sufficient condition for the reliability representation based on signature.

Systems with non-exchangeable components: Navarro et al. [56] studied measures like reliability and hazard rate by using averaging system and projection system. Navarro et al. [53] discussed the reliability function representation and stochstic orderings of system lifetimes when the components are independent and dependent. Cerqueti and Spizzichino [10] reviewed signature-related problems in the case of exchangeable and non-exchangeable components. In the cases when the components are independent and dependent, Samaniego and Navarro [62] discussed lifetime comparisons of systems of different sizes based on survival signature and distortion distribution, respectively. Navarro and Fernandez-Sanchez [47] studied signature-based reliability representations for coherent systems with identically distributed

dependent non-exchangeable components. Ding et al. [19] considered stochastic comparisons of heterogeneous systems with same size or different sizes by using survival signatures.

2.3 Computational Methods for Signatures

Signature concepts enjoy many useful properties and numerous applications, but that is due to the fact that they can be calculated efficiently. There are many different methods of calculating system signature, and each of them possess its own advantages and disadvantages.

Definition method [29]: Consider all component failure orderings and number of components in each ordering that causes the failure of the system; then, the structural signature can be obtained. This method is simple and direct, and does not need any other theory other than the definition itself. However, it needs a large amount of calculations and takes a long time, and would take considerable amount of time, and is therefore not suitable for large-scale complex systems.

Path/cut set method [8, 16]: Obtain all (minimal) path/cut sets of the system first, and then calculate the signature based on them. This method is more efficient for systems whose path/cut sets are known or easy to obtain, since the process of calculating path/cut sets itself would be time-assuming.

Reliability method [9, 31, 38, 60]: Calculate reliability polynomial of the system first, and then get signature from the polynomial. There are many effective calculation methods for reliability polynomials, but every method has its own limitations.

Binary decision graph method [57, 58]: Calculate the system signature based on binary decision graph of a network system. This method does not require the (minimal) path/cut sets, but it needs a considerable amount of storage and is therefore not suitable for large-scale systems.

Generating function method [63]: Obtain the signature of a consecutive-k-out-of-n system based on the generating function. This method gives explicit recursive formula for the system signature, but it is applicable only for consecutive type systems.

Five-component system algorithm[48]: Generate minimal cut sets of all coherent systems with five components by an algorithm, and then calculate minimal signatures and signatures for all of them. This algorithm gives signatures of all coherent systems with five components, but is not applicable for systems with more than five components.

Markov process method [64]: Derive the relationship between signature and the numbers of non-cut sets of different sizes based on Markov process, and then use it to obtain the signature. This method yields explicit formulas for signatures of some

common consecutive type systems, but it is not always easy to calculate the numbers of non-cut sets.

Module decomposition method [13, 15, 21, 23, 26, 28, 40]: Calculate the signature through the signature of each module and the signature of the structure. This method is suitable for large and complex systems and can be used in combination with other methods, but is only suitable for systems with modular characteristics.

2.4 Multi-state Signatures for Multi-state Systems

For multi-state systems in the field of reliability and network, some multi-state signature concepts have been proposed (related definitions are shown in Table 2). Levitin et al. [32] first defined two-dimensional D-spectrum for network systems, and then extended it to multi-dimensional D-spectrum. Subsequently, Gertsbakh and Shpungin [25] named marginal spectrums of the two-dimensional D-spectrum as the first spectrum and the second spectrum, respectively, and also named the corresponding cumulative spectrums as the first cumulative spectrum and the second cumulative spectrum. More related research on multi-dimensional D-spectrum in the field of network can be found in Lisnianski et al. [34].

Gertsbakh et al. [27] extended the concept of two-dimensional D-spectrum from the network field to the reliability field, where it is referred to as bivariate signature, and they also proposed bivariate tail signature based on the definition of tail signature for binary-state systems. Da and Hu [14] gave an equivalent definition of bivariate signature from probability viewpoint, and used it to study reliability representation of multi-state systems. Later, Ashrafi and Asadi [3] added a requirement of $i < j$ in the definition of bivariate signature, and called the new concept two-dimensional signature. Based on that definition, Zarezadeh et al. [75] further gave the definition of m-dimensional signature.

In analogy to the dynamic signature and conditional signature for binary-state systems, Ashrafi and Asadi [2] proposed the concept of dynamic signature matrix to discuss residual lifetimes of multi-state systems. Based on a decomposition of the structure functions of multi-state systems, Marichal et al. [39] studied two-dimensional signature and bivariate tail signature of multi-state systems by using the joint signature for binary-state systems.

The above indices are all for multi-state systems with binary-state components, and for multi-state systems with multi-state components, De Costa Bueno [17] defined multi-state signature for the first time, but the exchangebility assumption made in it for component lifetimes in all states may not be realistic in practice. Subsequently, Eryilmaz and Tuncel [22] proposed multi-state survival signature at system state level j for multi-state systems with multi-state components. Liu et al. [35] illustrated an application of this concept to stress-strength reliability systems. These laid the foundations for the signature theory for multi-state systems, but com-

Table 2 Definitions for multi-state signature concepts.

Field	Concept	Definition
Reliability	Bivariate signature	A matrix of proportions of component failure orderings such that the lth component failure causes the failure of the system partially and the rth component failure causes the failure of the system completely [27]
	Bivariate tail signature	A matrix of probabilities such that the system still works at time of the ith component failure and is not completely failed at time of the jth component failure [27]
	Two-dimensional signature	A matrix of probabilities such that the ith component failure causes the failure of the system partially and the jth component failure causes the failure of the system completely [3, 14]
	M-dimensional signature	A direct generalization of two-dimensional signature [75]
	Dynamic signature matrix	A matrix of conditional probabilities such that ith component failure causes the failure of the system partially and the jth one causes it completely, given the number of failed components [2]
	Multi-state signature	A matrix of probabilities such that the ith component that enters states below j causes the system to enter states below k [17]
	Survival signature	A collection of proportions of cut sets in all component state sets such that there are s_i components in state i or above for $i = 1, \ldots, j$ [22]
Network	Two-dimensional D-Spectrum	A collection of proportions of edge failure orderings such that a network falls apart into two disconnected clusters at the ith edge failure and then falls apart into three clusters at the jth edge failure [32]
	Multi-dimensional D-Spectrum	A direct generalization of two-dimensional D-Spectrum [32]
	First/second Spectrum	A collection of proportions such that a network falls apart into two/ three disconnected clusters at the ith edge failure [25]
	First/second cumulative Spectrum	A collection of proportions such that a network falls apart into two/ three disconnected clusters at the ith edge failure or before [25]

pared to signature theory for binary-state systems, there still remains a lot of work to be done on this topic.

In the following sections, we provide a detailed review of multi-state signature concepts in Sect. 3, and then various properties of these concepts in Sect. 4. In Sect. 5, we describe comparisons of multi-state systems of different sizes, and in Sect. 6, we discuss signature concepts of multi-state systems based on a given structure of modules. Finally, several illustrative examples are presented in Sect. 7, and some concluding remarks are made in Sect. 8.

3 Multi-state Signature Concepts

There are mainly two types of multi-state systems—multi-state systems with binary-state components and multi-state systems with multi-state components. Multi-state signature concepts can be defined for both these types, but the former is relatively easier to deal with. For example, considering a multi-state coherent system with state space $\{0, \ldots, n\}$ and m binary-state components with *i.i.d.* lifetimes $X_1, \ldots, X_m$

from a common absolutely continuous distribution function $F(x), x \geq 0$, the multi-state signature can be defined as follows to describe the structure of the system.

Definition 3.1 *(Multi-state signature)* The multi-state signature of the multi-state coherent system with m binary-state components is

$$\boldsymbol{s} = (s_{i_1,\ldots,i_n}, 1 \leq i_1 \leq \cdots \leq i_n \leq m),$$

where $s_{i_1,\ldots,i_n} = \mathbb{P}\{T_1 = X_{i_1:m}, \ldots, T_n = X_{i_n:m}\}$, $T_1, \ldots, T_n$ are the system lifetimes in state levels $n, \ldots, 1$, and $X_{i_1:m}, \ldots, X_{i_n:m}$ are the i_1th, $\ldots$, i_nth order statistics among the component lifetimes $X_1, \ldots, X_m$, respectively. See Yi et al. [65] for pertinent details.

Similarly, considering two multi-state coherent systems with the same state space $\{0, \ldots, n\}$ and m shared binary-state components with *i.i.d.* lifetimes $X_1, \ldots, X_m$ from a common absolutely continuous distribution function $F(x), x \geq 0$, the multi-state joint signature can be defined as follows to decribe the structures of the two systems.

Definition 3.2 *(Multi-state joint signature)* The multi-state joint signature of two multi-state coherent systems with m shared binary-state components is defined as

$$\boldsymbol{S} = (s_{i_1,\ldots,i_n;j_1,\ldots,j_n}, 1 \leq i_1 \leq \cdots \leq i_n \leq m, 1 \leq j_1 \leq \cdots \leq j_n \leq m),$$

where

$$s_{i_1,\ldots,i_n;j_1,\ldots,j_n} = \mathbb{P}\{T_1^{(1)} = X_{i_1:m}, \ldots, T_n^{(1)} = X_{i_n:m}, T_1^{(2)} = X_{j_1:m}, \ldots, T_n^{(2)} = X_{j_n:m}\},$$

and $T_1^{(1)}, \ldots, T_n^{(1)}, T_1^{(2)}, \ldots, T_n^{(2)}$ are the lifetimes on states $n, \ldots, 1$ of the two systems, $X_{i_1:m}, \ldots, X_{i_n:m}, X_{j_1:m}, \ldots, X_{j_n:m}$ are the i_1th, $\ldots$, i_nth, j_1th, $\ldots$, j_nth order statistics among the component lifetimes $X_1, \ldots, X_m$, respectively. See Yi et al. [67] for details.

Considering a multi-state coherent system with state space $\{0, \ldots, n\}$ and m binary-state components with *i.i.d.* lifetimes $X_1, \ldots, X_m$ from a common absolutely continuous distribution function $F(x), x \geq 0$, given that the system is in state k at time t with exactly i failed components, the dynamic multi-state signature can be defined as follows to describe the structure of the remaining system.

Definition 3.3 *(Dynamic multi-state signature)* Let $\boldsymbol{s} = (s_{i_1,\ldots,i_n}, 1 \leq i_1 \leq \cdots \leq i_n \leq m)$ be the multi-state signature of the orginal multi-state coherent system with m binary-state components. Then, the dynamic multi-state of the system is given by

$$\boldsymbol{s}^{(k)}(m-i) = \left(s^{(k)}_{i_{n-k+1},\ldots,i_n}(m-i), i+1 \leq i_{n-k+1} \leq \cdots \leq i_n \leq m\right),$$

where

$$s^{(k)}_{i_{n-k+1},\ldots,i_n}(m-i) = P\left\{T_{n-k+1} = X_{i_{n-k+1}:m}, \ldots, T_n = X_{i_n:m} \mid T_{n-k} \le t < T_{n-k+1}, X_{i:m} \le t < X_{i+1:m}\right\}$$

is the conditional probability that the system enters states $k-1, \ldots, 0$ at the i_{n-k+1}th, $\ldots, i_n$th ordered component failures, respectively, given that it is in state k at time t with exactly i failed components. Note that $k = 1, \ldots, n$, $i = 0, \ldots, m-1$, $T_0 = 0$, $X_{0:n} = 0$, and the component/system lifetimes are denoted in the same way as in Definition 3.1. See Yi et al. [72] for details.

Considering l independent and identical multi-state coherent systems in a life test, of which system p has state space $\{0, \ldots, n\}$ and m binary-state components with *i.i.d.* lifetimes $X^p_1, \ldots, X^p_m$ from a common absolutely continuous distribution function $F(x)$, $x \ge 0$, the ordered multi-state signature can be defined as follows to describe the structure of the qth system that enters state $n-r$ or below.

Definition 3.4 *(Ordered multi-state signature)* The ordered multi-state system signature for l multi-state coherent systems is defined as

$$\boldsymbol{s}^{qr} = (s^{qr}_{j_1,\ldots,j_n}, 1 \le j_1 \le \cdots \le j_n \le m),$$

for $q = 1, 2, \ldots, l$, $r = 1, 2, \ldots, n$, where

$$s^{qr}_{j_1,\ldots,j_n} = \mathbb{P}\{T^p_1 = X^p_{j_1:m}, \ldots, T^p_n = X^p_{j_n:m} \,|T^{q:l}_r = T^p_r\},$$

$T^{q:l}_r$ is the qth order statistic of lifetimes $T^1_r, \ldots, T^l_r$ for systems $1, \ldots, l$ at state level $n-r+1$, and $X^p_{j_1:m}, \ldots, X^p_{j_n:m}$ are the j_1th, $\ldots$, j_nth order statistics of the component lifetimes $X^p_1, \ldots, X^p_m$ for system p, respectively. See Yi et al. [65] for related details and see Yi et al. [72] for how this definition can be generalized to the case of non-identical multi-state coherent systems.

The above definitions are all for multi-state systems with binary-state components. Now, considering a multi-state coherent system with n *i.i.d.* multi-state components and a state space $\{0, \ldots, M\}$ for both the system and the components, the multi-state survival signature of the system can be defined as follows to describe the structure of the system.

Definition 3.5 *(Multi-state survival signature)* The multi-state survival signature of the multi-state coherent system is given by $\boldsymbol{S} = (\boldsymbol{S}^{(0)}, \ldots, \boldsymbol{S}^{(M)})$, where

$$\boldsymbol{S}^{(j)} = \left(S^{(j)}_{i_1,\ldots,i_M}, 0 \le i_M \le \ldots \le i_1 \le n\right), \; j = 0, \ldots, M,$$

is the multi-state survival signature at system state level j with

$$S^{(j)}_{i_1,\ldots,i_M} = P\left\{L_j > t \,|m_1(t) = i_1, \ldots, m_M(t) = i_M\right\}$$

being the conditional probability that the system is in state j or above at time t, given that $m_l(t) = i_l$ components are in state l or above, for all $l = 1, \ldots, M$. Here, $L_1, \ldots, L_M$ are the system lifetimes at state levels $1, \ldots, M$. See Yi et al. [73] and Yi et al. [71] for details.

Note that not all assumptions for these multi-state signature concepts are necessary for further work; for example, the systems can be semi-coherent instead of coherent, and the common component lifetime distribution can be continuous instead of absolutely continuous since its probability density funtion is not needed.

4 Properties of Multi-state Signatures

In this section, some important properties of multi-state signature concepts will be discussed. These include multi-state signature, multi-state joint signature, dynamic multi-state signature, ordered multi-state signature and multi-state survival signature.

4.1 Multi-state Signature

Based on the multi-state signature of a multi-state coherent system, the joint reliability function of system lifetimes at different state levels can be represented as follows.

Theorem 4.1 *For* $0 \le t_1 \le \cdots \le t_n \le \infty$, *we have*

$$\begin{aligned}
&\mathbb{P}\{T_1 > t_1, \ldots, T_n > t_n\} \\
&= \sum_{1 \le i_1 \le \cdots \le i_n \le m} s_{i_1,\ldots,i_n} \sum_{\substack{0 \le j_1 \le i_1 - 1, j_1 \le j_2 \le i_2 - 1, \\ \ldots, j_{n-1} \le j_n \le i_n - 1}} \left[\frac{m!}{j_1! \prod_{l=1}^{n-1} (j_{l+1} - j_l)!(m - j_n)!} \right] \\
&\quad \times F^{j_1}(t_1)[F(t_2) - F(t_1)]^{j_2 - j_1} \times \cdots \times [F(t_n) - F(t_{n-1})]^{j_n - j_{n-1}} F^{m - j_n}(t_n).
\end{aligned}$$

Proof See Theorem 2.1 in Yi et al. [65].

Another important property of multi-state signature is that its stochastic ordering, as defined in Yi et al. [65] will lead to stochastic ordering of the system lifetimes.

Theorem 4.2 *Suppose* $\boldsymbol{s}$ *and* $\tilde{\boldsymbol{s}}$ *are signatures of two multi-state systems with the same state space* $\{0, \ldots, n\}$ *and* m *binary-state components i.i.d. from the same absolutely continuous distribution function* $F(x), x \ge 0$. *If* $\boldsymbol{s} \le^{\text{st}} \tilde{\boldsymbol{s}}$, *then the lifetime vectors of the two systems are such that* $(T_1, \ldots, T_n) \le^{\text{st}} (\tilde{T}_1, \ldots, \tilde{T}_n)$.

Proof See Theorem 2.2 in Yi et al. [65].

4.2 Multi-state Joint Signature

Based on the multi-state joint signature of two multi-state coherent systems, the joint distribution function of the lifetimes of two systems at different state levels can be represented as follows.

Theorem 4.3 *For* $0 < t_1^{(1)} \le \cdots \le t_n^{(1)} < \infty$ *and* $0 < t_1^{(2)} \le \cdots \le t_n^{(2)} < \infty$, *the joint distribution function*

$$\begin{aligned} &G(t_1^{(1)}, \ldots, t_n^{(1)}, t_1^{(2)}, \ldots, t_n^{(2)}) \\ &= \mathbb{P}\{T_1^{(1)} \le t_1^{(1)}, \ldots, T_n^{(1)} \le t_n^{(1)}, T_1^{(2)} \le t_1^{(2)}, \ldots, T_n^{(2)} \le t_n^{(2)}\} \end{aligned}$$

of the system lifetimes $(T_1^{(1)}, \ldots, T_n^{(1)}, T_1^{(2)}, \ldots, T_n^{(2)})$ *can be expressed as*

$$\begin{aligned} &G(t_1^{(1)}, \ldots, t_n^{(1)}, t_1^{(2)}, \ldots, t_n^{(2)}) \\ &= \sum_{\substack{1 \le i_1 \le \cdots \le i_n \le m, \\ 1 \le j_1 \le \cdots \le j_n \le m}} s_{i_1,\ldots,i_n;j_1,\ldots,j_n} F_{i_1,\ldots,i_n;j_1,\ldots,j_n:m}(t_1^{(1)}, \ldots, t_n^{(1)}, t_1^{(2)}, \ldots, t_n^{(2)}), \end{aligned}$$

based on the multi-state joint signature $\boldsymbol{S}$ *given in Definition 3.2, where*

$$\begin{aligned} &F_{i_1,\ldots,i_n;j_1,\ldots,j_n:m}(t_1^{(1)}, \ldots, t_n^{(1)}, t_1^{(2)}, \ldots, t_n^{(2)}) \\ &= \mathbb{P}\{X_{i_1:m} \le t_1^{(1)}, \ldots, X_{i_n:m} \le t_n^{(1)}, X_{j_1:m} \le t_1^{(2)}, \ldots, X_{j_n:m} \le t_n^{(2)}\} \end{aligned}$$

is the joint distribution function of the order statistics $X_{i_1:m}, \ldots, X_{i_n:m}, X_{j_1:m}, \ldots, X_{j_n:m}$ *obtained from the i.i.d. component lifetimes* $X_1, \ldots, X_m$.

Proof See Theorem 2.2 in Yi et al. [67].

For two multi-state coherent systems with shared components, the comparison of their lifetimes is distribution-free and depends only on their joint signature, as described in the following theorem.

Theorem 4.4 *For two multi-state semi-coherent systems with a joint signature* $\boldsymbol{S}$, *their lifetimes are such that*

$$\begin{aligned} &\mathbb{P}(T_1^{(1)} \le T_1^{(2)}, \ldots, T_n^{(1)} \le T_n^{(2)}) \\ &= \sum_{1 \le i_1 \le \cdots \le i_n \le m, 1 \le j_1 \le \cdots \le j_n \le m, i_1 \le j_1, \ldots, i_n \le j_n} s_{i_1,\ldots,i_n;j_1,\ldots,j_n}. \end{aligned}$$

Proof See Sect. 3 in Yi et al. [67].

Stochastic comparisons of two independent pairs of dependent multi-state coherent systems can be presented based on the ordering of their multi-state joint signature.

Theorem 4.5 *For two multi-state joint signatures* $\boldsymbol{S}$ *and* $\tilde{\boldsymbol{S}}$*, if* $\boldsymbol{S} \leq_{S/E\rightarrow} \tilde{\boldsymbol{S}}$*, then their associated system lifetimes are such that*

$$(T_1^{(1)}, \ldots, T_n^{(1)}, T_1^{(2)}, \ldots, T_n^{(2)}) \leq_{\mathrm{st}} (\tilde{T}_1^{(1)}, \ldots, \tilde{T}_n^{(1)}, \tilde{T}_1^{(2)}, \ldots, \tilde{T}_n^{(2)}).$$

Here, $\boldsymbol{S} \leq_{S/E\rightarrow} \tilde{\boldsymbol{S}}$ *means that there exist matrices* $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ *of the same dimension and total mass such that* $\boldsymbol{S} = \boldsymbol{A}_1 \rightarrow \cdots \rightarrow \boldsymbol{A}_k = \tilde{\boldsymbol{S}}$*, where* $\boldsymbol{A}_l \rightarrow \boldsymbol{A}_{l+1}$ *means that* $\boldsymbol{A}_{l+1}$ *is obtained from* $\boldsymbol{A}_l$ *by moving a positive mass* $c_l > 0$ *from the* $(i_1, \ldots, i_n, j_1, \ldots, j_n)$*th term of* $\boldsymbol{A}_l$ *to its* $(\tilde{i}_1, \ldots, \tilde{i}_n, \tilde{j}_1, \ldots, \tilde{j}_n)$*th term with* $\tilde{i}_1 \geq i_1, \ldots, \tilde{i}_n \geq i_n$ *and* $\tilde{j}_1 \geq j_1, \ldots, \tilde{j}_n \geq j_n$*.*

Proof See Theorem 3.2 in Yi et al. [67].

4.3 Dynamic Multi-state Signature

The dynamic multi-state signature can be calculated directly from the corresponding multi-state signature and the number of failed components that cause the system to be in state k at a given time, as stated in the following theorem.

Theorem 4.6 *The dynamic multi-state system signature* $\boldsymbol{s}^{(k)}(m-i)$ *can be given from the multi-state signature* $\boldsymbol{s} = (s_{i_1,\ldots,i_n}, 1 \leq i_1 \leq \cdots \leq i_n \leq m)$ *as*

$$s^{(k)}_{i_{n-k+1},\ldots,i_n}(m-i) = \left(\sum_{\substack{1 \leq i_1 \leq \cdots \leq i_{n-k} \leq i < \\ i_{n-k+1} \leq \cdots \leq i_n \leq m}} s_{i_1,\ldots,i_n} \right)^{-1} \sum_{1 \leq i_1 \leq \cdots \leq i_{n-k} \leq i} s_{i_1,\ldots,i_n},$$

where $i + 1 \leq i_{n-k+1} \leq \cdots \leq i_n \leq m$*.*

Proof See Theorem 3.1 in Yi et al. [72].

4.4 Ordered Multi-state Signature

For independent and indentcal multi-state coherent systems in a life test, the ordered multi-state signature can be calculated directly from their common multi-state signature, as described in the following theorem.

Theorem 4.7 *Given the multi-state signature* $\boldsymbol{s} = (s_{i_1,\ldots,i_n}, 1 \leq i_1 \leq \cdots \leq i_n \leq m)$*, the multi-state ordered signture can be given as* $\boldsymbol{s}^{qr} = (s^{qr}_{j_1,\ldots,j_n}, 1 \leq j_1 \leq \cdots \leq j_n \leq m)$*, where*

$$s^{qr}_{j_1,\ldots,j_n} = \sum_{\boldsymbol{l}\in\mathcal{L}_{l,m}} \binom{l}{l_{i_1,\ldots,i_n}, 1 \le i_1 \le \cdots \le i_n \le m} \left\{ \prod_{1\le i_1\le\cdots\le i_n\le m} s^{l_{i_1,\ldots,i_n}}_{i_1,\ldots,i_n} \right\} p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}},$$

with $p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}}$ *being the conditional probability that the* q*th system that enters state* $n-r$ *or below enters states* $n-1,\ldots,0$ *or below at the* j_1*th, …,* j_n*th component failures, respectively, given* $\boldsymbol{l}$ *in the subset*

$$\mathcal{L}_{l,m} = \left\{ \boldsymbol{l} = (l_{i_1,\ldots,i_n}, 1 \le i_1 \le \cdots \le i_n \le m) : \sum_{1\le i_1\le\cdots\le i_n\le m} l_{i_1,\ldots,i_n} = l \right\}.$$

Proof See Theorem 3.1 in Yi et al. [65].

From the formula presented above, we can also readily conclude that the ordered multi-state signature is distribution-free.

Theorem 4.8 *The ordered multi-state system signature* $\boldsymbol{s}^{qr}$ *is free of the underlying component lifetime distribution function* F*, and is thus a distribution-free measure.*

Proof See Theorem 3.1 in Yi et al. [65].

The main difficulty in calculating the ordered multi-state signature is in the calculation of the required conditional probabilities $p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}}$. The following properties of the conditional probabilities will help reduce this computational difficulty.

Theorem 4.9 *The conditional probabilities* $p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}}$ *satisfy the following properties:* (1) *If* $l_{j_1,\ldots,j_n} = l$*, then* $p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}} = 1$*;*
(2) *If* $l_{j_1,\ldots,j_n} = 0$*, then* $p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}} = 0$*;*
(3) $\sum_{q=1}^{l} p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}} = l_{j_1,\ldots,j_n}$*;*
(4) $p^{qr}_{j_1,\ldots,j_n|\boldsymbol{l}} = p^{(l-q+1)(n-r+1)}_{m-j_n+1,\ldots,m-j_1+1|rev\ \boldsymbol{l}}$.

Proof See Theorem 3.2 in Yi et al. [65].

Theorem 4.10 *For any multi-state signature* $\boldsymbol{s}$ *and ordered multi-state signature* $\boldsymbol{s}^{qr}$ $(r = 1,\ldots,n, q = 1,\ldots,l)$*, all* $s^{qr}_{j_1,\ldots,j_n} = 0$ *if and only if* $s_{j_1,\ldots,j_n} = 0$*.*

Proof See Corollary 3.1 in Yi et al. [65].

Theorem 4.11 *The ordered signatures satisfy* $\frac{1}{l}\sum_{q=1}^{l} \boldsymbol{s}^{qr} = \boldsymbol{s}$ *and*

$$rev\ \boldsymbol{s}^{qr} = (rev\ \boldsymbol{s})^{(l-q+1)(n-r+1)}.$$

Proof See Theorem 3.3 in Yi et al. [65].

Theorem 4.12 *For any* $k \in \{1,\ldots,n\}$*, if* $s_{i_1,\ldots,i_n} = 0$ *for all* $1 \le i_1 \le \cdots \le i_n \le m$ *with* $i_k \ne a$*, then* $\boldsymbol{s}^{1k} = \cdots = \boldsymbol{s}^{lk} = \boldsymbol{s}$*.*

Proof See Theorem 3.4 in Yi et al. [65].

Ordered multi-state signatures $\boldsymbol{s}^{1r}, \ldots, \boldsymbol{s}^{lr}$ follow a specific ordering and they are equal to each other only under some special cases as described in the following theorem.

Theorem 4.13 (1) *For any* $1 \leq q_1 < q_2 \leq l$, *the ordered multi-state signatures satisfy* $\boldsymbol{s}^{q_1 r} \leq^{\mathrm{st}} \boldsymbol{s}^{q_2 r}$; (2) *If* $\boldsymbol{s}^{q_1 r} \geq^{\mathrm{st}} \boldsymbol{s}^{q_2 r}$ *for any* $1 \leq q_1 < q_2 \leq l$, *then there is at most one positive number in* $s_{j_1,\ldots,j_n}$ $(j_r = 1, \ldots, m)$ *for any* $1 \leq j_1 \leq \cdots \leq j_{r-1} \leq j_{r+1} \leq \cdots \leq j_n \leq m$.

Proof See Theorem 3.5 in Yi et al. [65].

Note that the properties discussed in this subsection can be generalized to independent and non-identical multi-state coherent systems; see Yi et al. [72] for related details and some further discussions.

4.5 Multi-state Survival Signature

Based on the multi-state survival signature of a multi-state coherent system with multi-state components, the reliabiltiy function of the system at each state level can be represented as follows.

Theorem 4.14 *The reliability of a multi-state system at level* k $(k = 0, \ldots, M)$ *is defined as the probability that the system is in state* k *or above at time* t. *Then, for a multi-state system with multi-state survival signature* $\boldsymbol{S} = (\boldsymbol{S}^{(0)}, \ldots, \boldsymbol{S}^{(M)})$ *and* n *i.i.d. multi-state components, the reliability function of the system at level* k *can be expressed as*

$$
\begin{aligned}
&R^{(k)}(t) = P\{T_k > t\} \\
&= \sum_{0 \leq i_1 + \cdots + i_M \leq n} S^{(k)}_{i_1 + \cdots + i_M, \ldots, i_M} \binom{n}{i_0, \ldots, i_M} [F_1(t)]^{i_0} \prod_{j=1}^{M-1} [F_{j+1}(t) - F_j(t)]^{i_j} [1 - F_M(t)]^{i_M},
\end{aligned}
$$

where $i_0 = n - \sum_{j=1}^{M} i_j$ *and* $F_j(t), t \geq 0$ $(j = 1, \ldots, M)$ *is the common absolutely continuous component lifetime distribution at state level* j.

Proof See Sect. 2 in Yi et al. [73].

Theorem 4.15 *Let* p_j $(j = 1, \ldots, M)$ *be the probability that each component is in state* j. *Then, the reliability at level* k *of a multi-state system with survival signature* $\boldsymbol{S} = (\boldsymbol{S}^{(0)}, \ldots, \boldsymbol{S}^{(M)})$ *can be given as*

$$
R^{(k)}(p_0, \ldots, p_M) = \sum_{0 \leq i_1 + \cdots + i_M \leq n} S^{(k)}_{i_1 + \cdots + i_M, \ldots, i_M} \binom{n}{i_0, \ldots, i_M} \prod_{j=0}^{M} p_j^{i_j}.
$$

Once reliability polynomials as above are determined for all level k, the multi-state survival signature of the system is also determined.

Proof See Sect. 2 in Yi et al. [73].

Note that the reliability polynomials provided above can be used to calculate multi-state survival signature based on reliability calculation methods. One may see Sect. 3 in Yi et al. [73] for related discussions based on the finite Markov chain imbedding approach.

5 Comparisons of Multi-state Systems of Different Sizes

Multi-state systems, no matter whether with binary-state components or with multi-state components, can be compared with each other througth their multi-state signature concepts. Multi-state signature concepts are always defined as matrices whose dimensions are determined by the numbers of components in the multi-state systems. Even though we have shown that stochastic orderings of these multi-state signatures lead to stochastic comparisons of associated multi-state systems, for multi-state systems of different sizes, transformation formulas are still needed for their stochastic comparsions, which is what is dicussed in this section.

5.1 Multi-state Systems of Sizes m and m + 1

In this subsection, transformation formulas of multi-state signature, multi-state joint signature and multi-state survival signature are presented for multi-state systems of sizes m and $m+1$ ($m \geq 2$).

Theorem 5.1 (Multi-state signature) *Let* $\boldsymbol{s} = (s_{k_1,\ldots,k_n}, 1 \leq k_1 \leq \cdots \leq k_n \leq m)$ *be the multi-state signature of a multi-state coherent or mixed system consisting of m binary-state components with i.i.d. continuous lifetimes. Then, its equivalent system with* $m+1$ *components has its multi-state signature as*

$$\boldsymbol{s}^* = \sum_{1 \leq k_1 \leq \cdots \leq k_n \leq m} s_{k_1,\ldots,k_n} \boldsymbol{s}^*_{k_1,\ldots,k_n:m},$$

where

$$\begin{aligned}\boldsymbol{s}^*_{k_1,\ldots,k_n:m} =& \frac{k_1}{m+1}\boldsymbol{s}_{k_1+1,\ldots,k_n+1:m+1} + \sum_{i=1}^{n-1} \frac{k_{i+1}-k_i}{m+1}\boldsymbol{s}_{k_1,\ldots,k_i,k_{i+1}+1,\ldots,k_n+1:m+1} \\ &+ \frac{m+1-k_n}{m+1}\boldsymbol{s}_{k_1,\ldots,k_n:m+1}.\end{aligned}$$

Here, $\boldsymbol{s}_{k_1,\ldots,k_n:m+1}$ is the multi-state signature of a multi-state $(k_n,\ldots,k_1)$-out-of-$m+1$: F system, and evidently, there is only one positive element in $\boldsymbol{s}_{k_1,\ldots,k_n:m+1}$.

Proof See Theorem 2.2 in Yi et al. [66].

Theorem 5.2 (Multi-state joint signature) *Let $\boldsymbol{s}=(s_{k_1,\ldots,k_n;r_1,\ldots,r_n}, 1\le k_1\le\cdots\le k_n\le m, 1\le r_1\le\cdots\le r_n\le m)$ be the multi-state joint signature of two multi-state semi-coherent or mixed systems consisting of m binary-state components with i.i.d. continuous lifetimes. Then, an equivalent pair of systems with $m+1$ binary-state components has their joint signature as*

$$\boldsymbol{s}^*=(s^*_{k_1,\ldots,k_n;r_1,\ldots,r_n}, 1\le k_1\le\cdots\le k_n\le m+1, 1\le r_1\le\cdots\le r_n\le m+1),$$

where

$$\begin{aligned}
&s^*_{k_1,\ldots,k_n;r_1,\ldots,r_n}\\
&=\frac{k_1\wedge r_1-1}{m+1}s_{k_1-1,\ldots,k_n-1;r_1-1,\ldots,r_n-1}+\frac{m+1-k_n\vee r_n}{m+1}s_{k_1,\ldots,k_n;r_1,\ldots,r_n}\\
&\quad+\sum_{b=0}^{n}\sum_{a=w_b+1}^{w_{b+1}-1}\frac{r_{a+1}-r_a-1}{m+1}s_{k_1,\ldots,k_b,k_{b+1}-1,\ldots,k_n-1;r_1,\ldots,r_a,r_{a+1}-1,\ldots,r_n-1}I_{\{w_b+1<w_{b+1},r_a<r_{a+1}\}}\\
&\quad+\sum_{b=1}^{n}\frac{k_b-r_{w_b}-1}{m+1}s_{k_1,\ldots,k_{b-1},k_b-1,\ldots,k_n-1;r_1,\ldots,r_{w_b},r_{w_b+1}-1,\ldots,r_n-1}I_{\{w_{b-1}<w_b,r_{w_b}<k_b\}}\\
&\quad+\sum_{b=1}^{n}\frac{r_{w_b+1}-k_b-1}{m+1}s_{k_1,\ldots,k_b,k_{b+1}-1,\ldots,k_n-1;r_1,\ldots,r_{w_b},r_{w_b+1}-1,\ldots,r_n-1}I_{\{w_b<w_{b+1},k_b<r_{w_b+1}\}}\\
&\quad+\sum_{b=1}^{n-1}\frac{k_{b+1}-k_b-1}{m+1}s_{k_1,\ldots,k_b,k_{b+1}-1,\ldots,k_n-1;r_1,\ldots,r_{w_b},r_{w_b+1}-1,\ldots,r_n-1}I_{\{w_b=w_{b+1},k_b<k_{b+1}\}},
\end{aligned}$$

with $k_0=r_0=1$, $k_{n+1}=r_{n+1}=m+1$ and w_j, $j=0,\ldots,n+1$, being the largest one in $\{i: r_i\le k_j, i=0,\ldots,n\}$.

Proof See Theorem 3.1 in Yi et al. [70].

Theorem 5.3 (Multi-state survival signature) *Let $\boldsymbol{S}=(\boldsymbol{S}^{(0)},\ldots,\boldsymbol{S}^{(M)})$, where $\boldsymbol{S}^{(i)}=\left(S^{(i)}_{i_1,\ldots,i_M}, 0\le i_M\le\cdots\le i_1\le n\right)$ $(i=0,\ldots,M)$, be the multi-state survival signature of a multi-state coherent or mixed system consisting of n i.i.d. multi-state components and a state space $\Omega=\{0,\ldots,M\}$ for both the system and the components. Suppose the component lifetimes $X_j^{(1)},\ldots,X_j^{(n)}$ $(j=1,\ldots,M)$ are i.i.d. with a common absolutely continuous distribution function $F_j(x)$, $x\ge 0$, and are independent for different j. Then, its equivalent system of size $n+1$ has its multi-state survival signature as*

$$\boldsymbol{S}^*=(\boldsymbol{S}^{(0)}_{n+1},\boldsymbol{S}^{*(1)},\ldots,\boldsymbol{S}^{*(M)})=\sum_{\boldsymbol{k}\in\mathcal{K}}s_{\boldsymbol{k}}\boldsymbol{S}^*_{\boldsymbol{k}:n},$$

where $\boldsymbol{S}^*_{\boldsymbol{k}:n} = (\boldsymbol{S}^{(0)}_{n+1}, \boldsymbol{S}^{*(1)}_{\boldsymbol{k}:n}, \ldots, \boldsymbol{S}^{*(M)}_{\boldsymbol{k}:n})$ *is the multi-state survival signature of the equivalent system of size* $n+1$ *of a multi-state* $\boldsymbol{k}$*-out-of-*n*: G system given by*

$$\boldsymbol{S}^{*(l)}_{\boldsymbol{k}:n} = \boldsymbol{S}^{*(l)}_{\boldsymbol{k}_l:n} = \left(S^{*(l)}_{\boldsymbol{k}_l,i_1,\ldots,i_M}, 0 \le i_M \le \cdots \le i_1 \le n+1\right), l = 1, \ldots, M,$$

with

$$S^{*(l)}_{\boldsymbol{k}_l,i_1,\ldots,i_M} = (n+1)^{-M} \prod_{j=1}^{M} \left[k_{l,j} I_{\{i_l \ge k_{l,j}+1\}} + (n+1-k_{l,j}) I_{\{i_l \ge k_{l,j}\}}\right]$$

for $0 \le i_M \le \cdots \le i_1 \le n+1$*, and*

$$\begin{aligned}\mathcal{K} =& \{(k_{i,j}, i = 1, \ldots, M, j = 1, \ldots, M) : 0 \le k_{i,j} \le k_{\tilde{i},\tilde{j}} \le n \text{ for any} \\ & 1 \le i < \tilde{i} \le M, 1 \le \tilde{j} < j \le M\},\end{aligned}$$

and $s_{\boldsymbol{k}}, \boldsymbol{k} \in \mathcal{K}$*, can be given as a solution to the set of linear equations*

$$\sum_{\boldsymbol{k}_l = \tilde{\boldsymbol{k}}} s_{\boldsymbol{k}} = s^{(l)}_{\tilde{\boldsymbol{k}}}, \quad \tilde{\boldsymbol{k}} \in \tilde{\mathcal{K}} = \{(k_1, \ldots, k_M) : 0 \le k_M \le \cdots \le k_1 \le n\}, \quad l = 1, \ldots, M,$$

with $\boldsymbol{s}^{(l)} = \left(s^{(l)}_{\tilde{\boldsymbol{k}}}, \tilde{\boldsymbol{k}} \in \tilde{\mathcal{K}}\right) = \boldsymbol{M}^{-1}\boldsymbol{S}^{(l)}$ *and*

$$\boldsymbol{M} = \left(M_{i_1,\ldots,i_M;j_1,\ldots,j_M}, 0 \le i_M \le \cdots \le i_1 \le n, 0 \le j_M \le \cdots \le j_1 \le n\right)$$

being a matrix with all elements $M_{i_1,\ldots,i_M;j_1,\ldots,j_M} = I_{\{i_1 \ge j_1,\ldots,i_M \ge j_M\}}$*, for all* $i = 1, \ldots, M$*.*

Proof See Theorem 2.2 in Yi et al. [71].

5.2 Multi-state Systems of Sizes m and $m + l$

In this subsection, transformation formulas of multi-state signature, multi-state joint signature and multi-state survival signature are presented for multi-state systems of sizes m and $m + l$ ($m \ge 2, l \ge 1$).

Theorem 5.4 (Multi-state signature) *Let* $\boldsymbol{s} = (s_{k_1,\ldots,k_n}, 1 \le k_1 \le \cdots \le k_n \le m)$ *be the multi-state signature of a multi-state coherent or mixed system consisting of* m *binary-state components with i.i.d. continuous lifetimes. Then, its equivalent system with* $m + l$ *components has its multi-state signature as*

$$\boldsymbol{s}^{(l)*} = \sum_{1 \le k_1 \le \cdots \le k_n \le m} s_{k_1,\ldots,k_n} \boldsymbol{s}^{(l)*}_{k_1,\ldots,k_n:m},$$

where, with $k_0 = i_0 = 0, k_{n+1} = m+1, i_{n+1} = m+l+1$,

$$\boldsymbol{s}^{(l)*}_{k_1,\ldots,k_n:m} = \sum_{(i_1,\ldots,i_n)\in\Omega_{m,l}(k_1,\ldots,k_n)} \binom{m+l}{m}^{-1} \prod_{s=0}^{n} \binom{i_{s+1}-i_s-1}{k_{s+1}-k_s-1}^{I_{\{k_{s+1}>k_s\}}} \boldsymbol{s}_{i_1,\ldots,i_n:m+l}$$

is the multi-state signature of an equivalent system with $m+l$ *components for a multi-state* $(k_n,\ldots,k_1)$*-out-of-*m*: F system with*

$$\Omega_{m,l}(k_1,\ldots,k_n) = \{(i_1,\ldots,i_n) : 1 \le i_1 \le \cdots \le i_n \le m+l,\ k_{s+1}-k_s \le i_{s+1}-i_s \text{ and } I_{\{i_{s+1}>i_s\}} = I_{\{k_{s+1}>k_s\}} \text{ for } s = 0,\ldots,n\}.$$

Proof See Theorem 2.5 in Yi et al. [66].

Theorem 5.5 (Multi-state joint signature) *Let* $\boldsymbol{s} = (s_{k_1,\ldots,k_n;r_1,\ldots,r_n}, 1 \le k_1 \le \cdots \le k_n \le m, 1 \le r_1 \le \cdots \le r_n \le m)$ *be the joint signature of two multi-state semi-coherent or mixed systems consisting of* m *binary-state components with i.i.d. continuous lifetimes. Then, an equivalent pair of two multi-state systems with* $m+l$ *binary-state components have the multi-state joint signature as*

$$\boldsymbol{s}^{(l)*} = \sum_{1\le k_1\le\cdots\le k_n\le m, 1\le r_1\le\cdots\le r_n\le m} s_{k_1,\ldots,k_n;r_1,\ldots,r_n} \boldsymbol{s}^{(l)*}_{k_1,\ldots,k_n;r_1,\ldots,r_n:m},$$

where

$$\begin{aligned}
&\boldsymbol{s}^{(l)*}_{k_1,\ldots,k_n;r_1,\ldots,r_n:m} \\
&= \sum_{(i_1,\ldots,i_n;j_1,\ldots,j_n)\in\Omega_{m,l}(k_1,\ldots,k_n;r_1,\ldots,r_n)} \binom{m+l}{m}^{-1} \binom{i_1\wedge j_1-1}{k_1\wedge r_1-1} \binom{m+l-i_n\vee j_n}{m-k_n\vee r_n} \\
&\quad\times \prod_{b=0}^{n} \prod_{a=w_b+1}^{w_{b+1}-1} \binom{j_{a+1}-j_a-1}{r_{a+1}-r_a-1}^{I_{\{w_b+1<w_{b+1},r_a<r_{a+1}\}}} \prod_{b=1}^{n} \binom{i_b-j_{w_b}-1}{k_b-r_{w_b}-1}^{I_{\{w_{b-1}<w_b,r_{w_b}<k_b\}}} \\
&\quad\times \prod_{b=1}^{n} \binom{j_{w_b+1}-i_b-1}{r_{w_b+1}-k_b-1}^{I_{\{w_b<w_{b+1},k_b<r_{w_b+1}\}}} \prod_{b=1}^{n-1} \binom{i_{b+1}-i_b-1}{k_{b+1}-k_b-1}^{I_{\{w_b=w_{b+1},k_b<k_{b+1}\}}} \\
&\quad\cdot \boldsymbol{s}_{i_1,\ldots,i_n;j_1,\ldots,j_n:m+l}
\end{aligned}$$

is the multi-state joint signature of the equivalent pair of systems with $m+l$ *shared components for a* $(k_n,\ldots,k_1)$*-out-of-*m*: F system and a* $(r_n,\ldots,r_1)$*-out-of-*m*: F system with* m *shared components, and*

$$\begin{aligned}
&\Omega_{m,l}(k_1,\ldots,k_n;r_1,\ldots,r_n) \\
&= \{(k_1+c_1,\ldots,k_n+c_n;r_1+d_1,\ldots,r_n+d_n) : 0 \le d_{w_0+1} \le \cdots \le d_{w_1} \le c_1 \le \cdots \le \\
&\quad d_{w_{n-1}+1} \le \cdots \le d_{w_n} \le c_n \le d_{w_n+1} \le \cdots \le d_{w_{n+1}} \le l \text{ and } I_{\{k_a=k_b\}} = I_{\{k_a+c_a=k_b+c_b\}}, \\
&\quad I_{\{r_a=r_b\}} \ge I_{\{r_a+d_a=r_b+d_b\}}, I_{\{k_a=r_b\}} \ge I_{\{k_a+c_a=r_b+d_b\}} \text{ for all } 1 \le a,b \le n\}.
\end{aligned}$$

Proof See Theorem 3.2 in Yi et al. [70].

Theorem 5.6 (Multi-state survival signature) *Let* $\boldsymbol{S} = (\boldsymbol{S}^{(0)}, \ldots, \boldsymbol{S}^{(M)})$, *where* $\boldsymbol{S}^{(i)} = \left(S^{(i)}_{i_1,\ldots,i_M}, 0 \le i_M \le \cdots \le i_1 \le n\right)$ $(i = 0, \ldots, M)$, *be the multi-state survival signature of a multi-state coherent or mixed system consisting of n i.i.d. multi-state components and a state space* $\Omega = \{0, \ldots, M\}$ *for both the system and the components. Suppose the component lifetimes* $X_j^{(1)}, \ldots, X_j^{(n)}$ $(j = 1, \ldots, M)$ *are i.i.d. with a common absolutely continuous distribution function* $F_j(x)$, $x \ge 0$, *and are independent for different* j. *Then, its equivalent system of size* $n + l$ *has its multi-state survival signature as*

$$\boldsymbol{S}^{[l]*} = (\boldsymbol{S}^{[l]*(0)}, \ldots, \boldsymbol{S}^{[l]*(M)}) = \sum_{\boldsymbol{k} \in \mathcal{K}} s_{\boldsymbol{k}} \boldsymbol{S}^{[l]*}_{\boldsymbol{k}:n},$$

where $s_{\boldsymbol{k}}, \boldsymbol{k} \in \mathcal{K}$, *are as in Theorem 5.3 and*

$$\boldsymbol{S}^{[l]*}_{\boldsymbol{k}:n} = (\boldsymbol{S}^{(0)}_{n+l}, \boldsymbol{S}^{[l]*(1)}_{\boldsymbol{k}:n}, \ldots, \boldsymbol{S}^{[l]*(M)}_{\boldsymbol{k}:n})$$

is the multi-state survival signature of the equivalent system of size $n + l$ *of a multi-state* $\boldsymbol{k}$*-out-of-n: G system given by* $\boldsymbol{S}^{[l]*(i)}_{\boldsymbol{k}:n} = \left(S^{[l]*(i)}_{\boldsymbol{k};i_1,\ldots,i_M}, 0 \le i_M \le \cdots \le i_1 \le n + l\right)$ $(i = 1, \ldots, M)$ *with* r_j $(j = 1, \ldots, M)$ *being the number of zeros in* $k_{1,j}, \ldots, k_{M,j}$, *and*

$$S^{[l]*(i)}_{\boldsymbol{k};i_1,\ldots,i_M} = \sum_{\boldsymbol{h} \in \mathcal{H}_{\boldsymbol{k}}} \prod_{j=1}^{M} \left\{ (n+l)^{-1} \binom{h_{r_j+1,j} - 1}{k_{r_j+1,j} - 1} \left[\prod_{s=r_j+1}^{M-1} \binom{h_{s+1,j} - h_{s,j} - 1}{k_{s+1,j} - k_{s,j} - 1}^{I_{\{k_{s+1,j} > k_{s,j}\}}} \right] \right.$$
$$\left. \binom{n + l - h_{M,j}}{n - k_{M,j}}^{I_{\{r_j < M\}}} \right\} I_{\{i_1 \ge h_{i,1}, \ldots, i_M \ge h_{i,M}\}}, \quad 0 \le i_M \le \cdots \le i_1 \le n + l,$$

with

$$\begin{aligned} \mathcal{H}_{\boldsymbol{k}} = \{&(h_{i,j}, j = 1, \ldots, M, i = r_j + 1, \ldots, M) : 1 \le h_{r_j+1,j} \le \cdots \le h_{M,j} \le n + l, \\ &k_{r_j+1,j} \le h_{r_j+1,j}, k_{r_j+2,j} - k_{r_j+1,j} \le h_{r_j+2,j} - h_{r_j+1,j}, \ldots, k_{M,j} - k_{M-1,j} \\ &\le h_{M,j} - h_{M-1,j}, h_{M,j} \le k_{M,j} + l, I_{\{k_{r_j+2,j} > k_{r_j+1,j}\}} = I_{\{h_{r_j+2,j} > h_{r_j+1,j}\}}, \\ &\ldots, I_{\{k_{M,j} > k_{M-1,j}\}} = I_{\{h_{M,j} > h_{M-1,j}\}} \text{ for all } j\}. \end{aligned}$$

Proof See Theorem 2.4 in Yi et al. [71].

6 Multi-state Systems Based on a Structure of Modules

For the calculation of signature concepts, module decomposition method is a very effective tool since it can be used together with other methods, as long as the system

has a structure of modules. In this section, we present some results for multi-state signature and multi-state survival signature for a series/parallel/recurrent structure of modules.

6.1 Series Structure of Modules

For two multi-state systems consisting of binary-state components, multi-state signature can be used to describe their structures, and multi-state signature of their series connection can be obtained as given in the following theorem.

Theorem 6.1 (Multi-state signature) *For two independent multi-state coherent systems ϕ_1 and ϕ_2 with state space $\{0, \ldots, M\}$ consisting of binary-state components which are i.i.d. from a continuous distribution function $F(x), x \geq 0$, the multi-state signature of a new system formed by their series connection can be computed from their multi-state signatures $\boldsymbol{s}^1 = (s^1_{i_1,\ldots,i_M}, 1 \leq i_1 \leq \cdots \leq i_M \leq m_1)$ and $\boldsymbol{s}^2 = (s^2_{i_1,\ldots,i_M}, 1 \leq i_1 \leq \cdots \leq i_M \leq m_2)$ as*

$$\boldsymbol{s}^{(ser)} = (s_{v_1,\ldots,v_M}, 1 \leq v_1 \leq \cdots \leq v_M \leq m_1 + m_2),$$

where

$$s_{v_1,\ldots,v_M} = \sum_{v_1-1 \leq u_1 \leq v_1,\ldots,v_M-1 \leq u_M \leq v_M} (-1)^{(u_1+\cdots+u_M)-(v_1+\cdots+v_M)+M} \bar{S}_{v_1,\ldots,v_M}.$$

Here, for $0 \leq u_1 \leq \cdots \leq u_M \leq m_1 + m_2$,

$$\bar{S}_{u_1,\ldots,u_M} = \sum_{(i_1,\ldots,i_M) \in W_{u_1,\ldots,u_M}} \bar{S}^1_{i_1,\ldots,i_M} \bar{S}^2_{u_1-i_1,\ldots,u_M-i_M} \prod_{r=0}^{M} \binom{u_{r+1}-u_r}{i_{r+1}-i_r} \binom{m_1+m_2}{m_1}^{-1}$$

with $u_0 = i_0 = 0, u_{M+1} = m_1 + m_2, i_{M+1} = m_1$, and

$$\bar{S}^r_{i_1,\ldots,i_M} = \sum_{i_1 < j_1 \leq m_1,\ldots,i_M < j_M \leq m_1} s^r_{j_1,\ldots,j_M}, r = 1, 2,$$

$$W_{u_1,\ldots,u_M} = \{(i_1, \ldots, i_M) : 1 \leq i_1 \leq \cdots \leq i_M \leq m_1, 1 \leq u_1 - i_1 \leq \cdots \leq u_M - i_M \leq m_2\};$$

otherwise, $\bar{S}_{u_1,\ldots,u_M} = \bar{S}_{u_1,\max(u_1,u_2),\ldots,\max(u_1,\ldots,u_M)}$.

Proof See Theorem 2.1 in Yi et al. [69].

For two multi-state systems with multi-state components, multi-state survival signature can be used to describe their structures, and multi-state signature of their series connection can be obtained as follows.

Theorem 6.2 (Multi-state survival signature) *For two independent multi-state coherent systems ϕ_1 and ϕ_2 with state space $\{0, \ldots, M\}$ consisting of multi-state components that enters states below r $(r = 1, \ldots, M)$ at times i.i.d. from a continuous distribution function $F_r(x), x \geq 0$, the multi-state survival signature of a new system formed by their series connection can be computed from their multi-state survival signatures $\boldsymbol{D}^1 = (\boldsymbol{D}^{1(0)}, \ldots, \boldsymbol{D}^{1(M)})$ and $\boldsymbol{D}^2 = (\boldsymbol{D}^{2(0)}, \ldots, \boldsymbol{D}^{2(M)})$ as $\boldsymbol{D}^{(ser)} = (\boldsymbol{D}^{(0)}, \ldots, \boldsymbol{D}^{(M)})$, where, for $a = 0, \ldots, M$,*

$$\begin{aligned}\boldsymbol{D}^{r(a)} &= (D^{r(a)}_{i_1,\ldots,i_M}, 0 \leq i_M \leq \ldots \leq i_1 \leq m_1), r = 1, 2,\\ \boldsymbol{D}^{(a)} &= (D^{(a)}_{u_1,\ldots,u_M}, 0 \leq u_M \leq \ldots \leq u_1 \leq m_1 + m_2),\end{aligned}$$

and

$$D^{(a)}_{u_1,\ldots,u_M} = \sum_{(i_1,\ldots,i_M)\in V_{u_1,\ldots,u_M}} D^{1(a)}_{i_1,\ldots,i_M} D^{2(a)}_{u_1-i_1,\ldots,u_M-i_M} \prod_{r=0}^{M} \binom{u_r - u_{r+1}}{i_r - i_{r+1}} \binom{m_1 + m_2}{m_1}^{-1}$$

with $u_0 = m_1 + m_2, i_0 = m_1, u_{M+1} = i_{M+1} = 0$, and

$$V_{u_1,\ldots,u_M} = \{(i_1, \ldots, i_M) : 1 \leq i_M \leq \cdots \leq i_1 \leq m_1, 1 \leq u_M - i_M \leq \cdots \leq u_1 - i_1 \leq m_2\}.$$

Proof See Theorem 3.1 in Yi et al. [69].

6.2 Parallel Structure of Modules

As in the case of series structure of modules, for two multi-state systems consisting of binary-state components, multi-state signature of their parallel connection can be obtained as presented in the following theorem.

Theorem 6.3 (Multi-state signature) *For two independent multi-state coherent systems ϕ_1 and ϕ_2 with state space $\{0, \ldots, M\}$ consisting of binary-state components which are i.i.d. from a continuous distribution function $F(x), x \geq 0$, the multi-state signature of a new system formed by their parallel connection can be computed from their multi-state signatures $\boldsymbol{s}^1 = (s^1_{i_1,\ldots,i_M}, 1 \leq i_1 \leq \cdots \leq i_M \leq m_1)$ and $\boldsymbol{s}^2 = (s^2_{i_1,\ldots,i_M}, 1 \leq i_1 \leq \cdots \leq i_M \leq m_2)$ as*

$$\boldsymbol{s}^{(par)} = (s_{v_1,\ldots,v_M}, 1 \leq v_1 \leq \cdots \leq v_M \leq m_1 + m_2),$$

where

$$s_{v_1,\ldots,v_M} = \sum_{v_1-1\leq u_1\leq v_1,\ldots,v_M-1\leq u_M\leq v_M} (-1)^{(v_1+\cdots+v_M)-(u_1+\cdots+u_M)} S_{u_1,\ldots,u_M},$$

Here, for $0 \le u_1 \le \cdots \le u_M \le m_1 + m_2$,

$$S_{u_1,\ldots,u_M} = \sum_{(i_1,\ldots,i_M)\in W_{u_1,\ldots,u_M}} S^1_{i_1,\ldots,i_M} S^2_{u_1-i_1,\ldots,u_M-i_M} \prod_{r=0}^{M} \binom{u_{r+1}-u_r}{i_{r+1}-i_r} \binom{m_1+m_2}{m_1}^{-1},$$

with notations as used in Theorem 6.1 and

$$S^r_{i_1,\ldots,i_M} = \sum_{1\le j_1<i_1,\ldots,1\le j_M<i_M} s^r_{j_1,\ldots,j_M}, r = 1, 2;$$

otherwise, $S_{u_1,\ldots,u_M} = S_{\min(u_1,\ldots,u_M),\ldots,\min(u_{M-1},u_M),u_M}$.

Proof See Theorem 2.2 in Yi et al. [69].

In an analogous way, for two multi-state systems consisting of multi-state components, multi-state survival signature of their parallel connection can be obtained as given in the following theorem.

Theorem 6.4 (Multi-state survival signature) *For two independent multi-state coherent systems* ϕ_1 *and* ϕ_2 *with state space* $\{0, \ldots, M\}$ *consisting of multi-state components that enter states below* r $(r = 1, \ldots, M)$ *at times i.i.d. from a continuous distribution function* $F_r(x)$, $x \ge 0$, *the multi-state survival signature of a new system formed by their parallel connection can be computed from their multi-state survival signatures* $\boldsymbol{D}^1 = (\boldsymbol{D}^{1(0)}, \ldots, \boldsymbol{D}^{1(M)})$ *and* $\boldsymbol{D}^2 = (\boldsymbol{D}^{2(0)}, \ldots, \boldsymbol{D}^{2(M)})$ *as* $\boldsymbol{D}^{(par)} = (\boldsymbol{D}^{(0)}, \ldots, \boldsymbol{D}^{(M)})$, *where, for* $a = 0, \ldots, M$, $\boldsymbol{D}^{1(a)}$, $\boldsymbol{D}^{2(a)}$, $\boldsymbol{D}^{(a)}$ *are defined in the same way as in Theorem 6.2 and*

$$D^{(a)}_{u_1,\ldots,u_M} = \left\{ 1 - \sum_{(i_1,\ldots,i_M)\in V_{u_1,\ldots,u_M}} \left[1 - D^{1(a)}_{i_1,\ldots,i_M}\right] \left[1 - D^{2(a)}_{u_1-i_1,\ldots,u_M-i_M}\right] \times \prod_{r=0}^{M} \binom{u_r-u_{r+1}}{i_r-i_{r+1}} \binom{m_1+m_2}{m_1}^{-1} \right\},$$

with notations being the same as in Theorem 6.2.

Proof See Theorem 3.2 in Yi et al. [69].

6.3 Recurrent Structure of Modules

As with series/parallel structure of modules, for two multi-state systems consisting of binary-state components, multi-state signature of their recurrent connection can be obtained as given in the following theorem.

Theorem 6.5 (Multi-state signature) *For N independent and identical multi-state coherent systems with state space $\{0,\ldots,M\}$ consisting of binary-state components which are i.i.d. from a continuous distribution function $F(x)$, $x \geq 0$, the multi-state signature of a new system formed by their recurrent connection can be computed from their common multi-state signature $\boldsymbol{s}^1 = (s^1_{i_1,\ldots,i_M}, 1 \leq i_1 \leq \cdots \leq i_M \leq m)$ and multi-state survival signature $\boldsymbol{D}^2 = (\boldsymbol{D}^{2(0)},\ldots,\boldsymbol{D}^{2(M)})$ of the connection as $\boldsymbol{D}^{(rec)} = (\boldsymbol{D}^{(0)},\ldots,\boldsymbol{D}^{(M)})$, where, for $a = 0,\ldots,M$,*

$$\boldsymbol{D}^{2(a)} = (D^{2(a)}_{j_1,\ldots,j_M}, 0 \leq j_M \leq \cdots \leq j_1 \leq N), \quad \boldsymbol{D}^{(a)} = (D^{(a)}_0,\ldots,D^{(a)}_{Nm}),$$

and

$$D^{(a)}_u = \binom{Nm}{u}^{-1} \sum_{0\leq j_1+\cdots+j_M\leq N} D^{2(a)}_{j_1+\cdots+j_M,\ldots,j_M} \binom{N}{j_0,\ldots,j_M} \times \sum_{\boldsymbol{i}\in\Theta_{u,j_0,\ldots,j_M}} \prod_{r=0}^{M} \left\{ \binom{j_r}{i_{r0},\ldots,i_{rm}} \prod_{k=0}^{M} \left[\left(D^{1(r)}_k - D^{1(r+1)}_k \right) \binom{m}{k} \right]^{i_{rk}} \right\}$$

with $u = 0,\ldots,Nm$, $j_0 = N - \sum_{l=1}^{M} j_l$, $D^{1(r)}_k = \sum_{i_{M-r+1}>k, 1\leq i_1\leq\cdots\leq i_M\leq m} s^1_{i_1,\ldots,i_M}$ $(r = 1,\ldots,M)$, $D^{1(0)}_k = 1$, $D^{1(M+1)}_k = 0$, *and*

$$\Theta_{u,j_0,\ldots,j_M} = \left\{ \boldsymbol{i} = (i_{rk}, 0 \leq r \leq M, 0 \leq k \leq m) : i_{r0} + \cdots + i_{rm} = j_r \text{ for any } r, \text{ and } \sum_{r=0}^{M}\sum_{k=0}^{m} k i_{rk} = u \right\}.$$

Proof See Theorem 4.1 in Yi et al. [69].

In an analogous way, for two multi-state systems consisting of multi-state components, multi-state survival signature of their recurrent connection can be obtained as given in the following theorem.

Theorem 6.6 (Multi-state survival signature) *For N independent multi-state coherent systems with state space $\{0,\ldots,M\}$ consisting of multi-state components that enter states below r $(r = 1,\ldots,M)$ at times i.i.d. from a continuous distribution function $F_r(x)$, $x \geq 0$, the multi-state survival signature of a new system formed by their recurrent connection can be computed from their common multi-state survival signature $\boldsymbol{D}^1 = (\boldsymbol{D}^{1(0)},\ldots,\boldsymbol{D}^{1(M)})$ and $\boldsymbol{D}^2 = (\boldsymbol{D}^{2(0)},\ldots,\boldsymbol{D}^{2(M)})$ as $\boldsymbol{D}^{(rec)} = (\boldsymbol{D}^{(0)},\ldots,\boldsymbol{D}^{(M)})$, where for $a = 0,\ldots,M$, $\boldsymbol{D}^{1(a)}$, $\boldsymbol{D}^{2(a)}$, $\boldsymbol{D}^{(a)}$ are as defined in Theorem 6.2, and*

$$
\begin{aligned}
&D^{(a)}_{u_1+\cdots+u_M,\ldots,u_M} \\
&=\binom{Nm}{u_0,\ldots,u_M}^{-1} \sum_{0\le j_1+\cdots+j_M\le N} D^{2(a)}_{j_1+\cdots+j_M,\ldots,j_M} \binom{N}{j_0,\ldots,j_M} \sum_{\boldsymbol{I}\in\Omega_{u_0,\ldots,u_M,j_0,\ldots,j_M}} \prod_{r=0}^{M} \\
&\quad\left\{\binom{j_r}{l_{i_{r1}+\cdots+i_{rM},\ldots,i_{rM}}, 0\le i_{r1}+\cdots+i_{rM}\le m} \prod_{0\le i_{r1}+\cdots+i_{rM}\le m} \left[\left(D^{1(r)}_{i_{r1}+\cdots+i_{rM},\ldots,i_{rM}}\right.\right.\right. \\
&\quad\left.\left.\left.-D^{1(r+1)}_{i_{r1}+\cdots+i_{rM},\ldots,i_{rM}}\right)\binom{m}{i_{r0},\ldots,i_{rM}}\right]^{l_{i_{r1}+\cdots+i_{rM},\ldots,i_{rM}}}\right\} I_{\{0\le u_1+\cdots+u_M\le Nm\}}
\end{aligned}
$$

with $u_0 = Nm - \sum_{l=1}^{M} u_l$, $j_0 = N - \sum_{l=1}^{M} j_l$, $i_{r0} = m - \sum_{l=1}^{M} i_{rl}$, *and*

$$
\begin{aligned}
&\Omega_{u_0,\ldots,u_M,j_0,\ldots,j_M} \\
&=\left\{\boldsymbol{l}=(l_{i_{r1}+\cdots+i_{rM},\ldots,i_{rM}}, 0\le i_{r1}+\cdots+i_{rM}\le m): \sum_{0\le i_{s1}+\cdots+i_{sM}\le m} l_{i_{s1}+\cdots+i_{sM},\ldots,i_{sM}}\right. \\
&\quad\left.=j_s,\ \sum_{r=0}^{M}\sum_{0\le i_{r1}+\cdots+i_{rM}\le m} i_{rs} l_{i_{r1}+\cdots+i_{rM},\ldots,i_{rM}} = u_s \text{ for any } s\right\}.
\end{aligned}
$$

Proof See Theorem 4.2 in Yi et al. [69].

7 Illustrative Examples

For theoretical results presented in above sections, their illustrative examples can be found in related papers, for example, see Yi et al. [69] for illustratve examples for Sect. 6. In this section, we present several illustrative examples for the transformation formulas of multi-state signature, multi-state joint signature and multi-state survival signature.

Example 7.1 *(Multi-state signature)* From Theorem 5.1, for a ternary-state coherent or mixed system consisting of two *i.i.d.* components that has multi-state signature

$$
\boldsymbol{s} = \begin{pmatrix} s_{1,1} & s_{1,2} \\ 0 & s_{2,2} \end{pmatrix} = \begin{pmatrix} 0 & 1/2 \\ 0 & 1/2 \end{pmatrix},
$$

its equivalent system with three components has its multi-state signature as

$$
\boldsymbol{s}^* = \frac{1}{2}\boldsymbol{s}^*_{1,2:2} + \frac{1}{2}\boldsymbol{s}^*_{2,2:2} = \frac{1}{2}\begin{pmatrix} 0 & 1/3 & 1/3 \\ 0 & 0 & 1/3 \\ 0 & 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 2/3 \end{pmatrix} = \begin{pmatrix} 0 & 1/6 & 1/6 \\ 0 & 1/6 & 1/6 \\ 0 & 0 & 1/3 \end{pmatrix}.
$$

From Theorem 5.4, its equivalent system with three components has its multi-state signature as

$$\boldsymbol{s}^{(2)*} = \frac{1}{2}\boldsymbol{s}^{(2)*}_{1,2:2} + \frac{1}{2}\boldsymbol{s}^{(2)*}_{2,2:2} = \frac{1}{2}\begin{pmatrix} 0 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/6 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 0 & 1/12 & 1/12 & 1/12 \\ 0 & 1/12 & 1/12 & 1/12 \\ 0 & 0 & 1/6 & 1/12 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}.$$

The same result can also be obtained by using Theorem 5.1 twice; see Sect. 3 in Yi et al. [66] for details.

Example 7.2 *(Multi-state joint signature)* From Theorem 5.2, for two ternary-state coherent or mixed systems with two shared *i.i.d.* components that has multi-state joint signature

$$\boldsymbol{s} = \begin{pmatrix} s_{1,1;1,1} & s_{1,1;1,2} & s_{1,1;2,2} \\ s_{1,2;1,1} & s_{1,2;1,2} & s_{1,2;2,2} \\ s_{2,2;1,1} & s_{2,2;1,2} & s_{2,2;2,2} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

their equivalent pair of systems with three *i.i.d.* shared components has their multi-state joint signature as

$$\boldsymbol{s}^* = \begin{pmatrix} s^*_{1,1;1,1} & s^*_{1,1;1,2} & s^*_{1,1;1,3} & s^*_{1,1;2,2} & s^*_{1,1;2,3} & s^*_{1,1;3,3} \\ s^*_{1,2;1,1} & s^*_{1,2;1,2} & s^*_{1,2;1,3} & s^*_{1,2;2,2} & s^*_{1,2;2,3} & s^*_{1,2;3,3} \\ s^*_{1,3;1,1} & s^*_{1,3;1,2} & s^*_{1,3;1,3} & s^*_{1,3;2,2} & s^*_{1,3;2,3} & s^*_{1,3;3,3} \\ s^*_{2,2;1,1} & s^*_{2,2;1,2} & s^*_{2,2;1,3} & s^*_{2,2;2,2} & s^*_{2,2;2,3} & s^*_{2,2;3,3} \\ s^*_{2,3;1,1} & s^*_{2,3;1,2} & s^*_{2,3;1,3} & s^*_{2,3;2,2} & s^*_{2,3;2,3} & s^*_{2,3;3,3} \\ s^*_{3,3;1,1} & s^*_{3,3;1,2} & s^*_{3,3;1,3} & s^*_{3,3;2,2} & s^*_{3,3;2,3} & s^*_{3,3;3,3} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

From Theorem 5.5, their equivalent pair of systems with four *i.i.d.* shared components has their multi-state joint signature as

$$
\boldsymbol{s}^{(2)*} = \begin{pmatrix}
s^{(2)*}_{1,1;1,1} & s^{(2)*}_{1,1;1,2} & s^{(2)*}_{1,1;1,3} & s^{(2)*}_{1,1;1,4} & s^{(2)*}_{1,1;2,2} & s^{(2)*}_{1,1;2,3} & s^{(2)*}_{1,1;2,4} & s^{(2)*}_{1,1;3,3} & s^{(2)*}_{1,1;3,4} & s^{(2)*}_{1,1;4,4} \\
s^{(2)*}_{1,2;1,1} & s^{(2)*}_{1,2;1,2} & s^{(2)*}_{1,2;1,3} & s^{(2)*}_{1,2;1,4} & s^{(2)*}_{1,2;2,2} & s^{(2)*}_{1,2;2,3} & s^{(2)*}_{1,2;2,4} & s^{(2)*}_{1,2;3,3} & s^{(2)*}_{1,2;3,4} & s^{(2)*}_{1,2;4,4} \\
s^{(2)*}_{1,3;1,1} & s^{(2)*}_{1,3;1,2} & s^{(2)*}_{1,3;1,3} & s^{(2)*}_{1,3;1,4} & s^{(2)*}_{1,3;2,2} & s^{(2)*}_{1,3;2,3} & s^{(2)*}_{1,3;2,4} & s^{(2)*}_{1,3;3,3} & s^{(2)*}_{1,3;3,4} & s^{(2)*}_{1,3;4,4} \\
s^{(2)*}_{1,4;1,1} & s^{(2)*}_{1,4;1,2} & s^{(2)*}_{1,4;1,3} & s^{(2)*}_{1,4;1,4} & s^{(2)*}_{1,4;2,2} & s^{(2)*}_{1,4;2,3} & s^{(2)*}_{1,4;2,4} & s^{(2)*}_{1,4;3,3} & s^{(2)*}_{1,4;3,4} & s^{(2)*}_{1,4;4,4} \\
s^{(2)*}_{2,2;1,1} & s^{(2)*}_{2,2;1,2} & s^{(2)*}_{2,2;1,3} & s^{(2)*}_{2,2;1,4} & s^{(2)*}_{2,2;2,2} & s^{(2)*}_{2,2;2,3} & s^{(2)*}_{2,2;2,4} & s^{(2)*}_{2,2;3,3} & s^{(2)*}_{2,2;3,4} & s^{(2)*}_{2,2;4,4} \\
s^{(2)*}_{2,3;1,1} & s^{(2)*}_{2,3;1,2} & s^{(2)*}_{2,3;1,3} & s^{(2)*}_{2,3;1,4} & s^{(2)*}_{2,3;2,2} & s^{(2)*}_{2,3;2,3} & s^{(2)*}_{2,3;2,4} & s^{(2)*}_{2,3;3,3} & s^{(2)*}_{2,3;3,4} & s^{(2)*}_{2,3;4,4} \\
s^{(2)*}_{2,4;1,1} & s^{(2)*}_{2,4;1,2} & s^{(2)*}_{2,4;1,3} & s^{(2)*}_{2,4;1,4} & s^{(2)*}_{2,4;2,2} & s^{(2)*}_{2,4;2,3} & s^{(2)*}_{2,4;2,4} & s^{(2)*}_{2,4;3,3} & s^{(2)*}_{2,4;3,4} & s^{(2)*}_{2,4;4,4} \\
s^{(2)*}_{3,3;1,1} & s^{(2)*}_{3,3;1,2} & s^{(2)*}_{3,3;1,3} & s^{(2)*}_{3,3;1,4} & s^{(2)*}_{3,3;2,2} & s^{(2)*}_{3,3;2,3} & s^{(2)*}_{3,3;2,4} & s^{(2)*}_{3,3;3,3} & s^{(2)*}_{3,3;3,4} & s^{(2)*}_{3,3;4,4} \\
s^{(2)*}_{3,4;1,1} & s^{(2)*}_{3,4;1,2} & s^{(2)*}_{3,4;1,3} & s^{(2)*}_{3,4;1,4} & s^{(2)*}_{3,4;2,2} & s^{(2)*}_{3,4;2,3} & s^{(2)*}_{3,4;2,4} & s^{(2)*}_{3,4;3,3} & s^{(2)*}_{3,4;3,4} & s^{(2)*}_{3,4;4,4} \\
s^{(2)*}_{4,4;1,1} & s^{(2)*}_{4,4;1,2} & s^{(2)*}_{4,4;1,3} & s^{(2)*}_{4,4;1,4} & s^{(2)*}_{4,4;2,2} & s^{(2)*}_{4,4;2,3} & s^{(2)*}_{4,4;2,4} & s^{(2)*}_{4,4;3,3} & s^{(2)*}_{4,4;3,4} & s^{(2)*}_{4,4;4,4}
\end{pmatrix}
$$

$$
= \begin{pmatrix}
0 & 0 & 0 & 0 & 1/6 & 0 & 0 & 1/6 & 0 & 1/6 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 0 & 1/6 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
$$

The same result can also be obtained by using Theorem 5.2 twice; see Sect. 4 in Yi et al. [70] for details.

Example 7.3 *(Multi-state survival signature)* For a multi-state linear consecutive (2, 1)-out-of-4:G system, its multi-state survival signature can be given as

$$
\boldsymbol{S} = (\boldsymbol{S}^{(0)}, \boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)}) = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 1/2 & 1 & 1 & 0 & 1/2 & 1 & 1 & 1/2 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1 & 1 & 1/2 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}^T.
$$

From Theorem 5.3, the multi-state survival signature of its equivalent system of size 5 can be given as $\boldsymbol{S}^* = \frac{1}{2}\boldsymbol{S}^*_{2,0;2,1:4} + \frac{1}{2}\boldsymbol{S}^*_{3,0;3,1:4}$, where $\boldsymbol{S}^*_{2,0;2,1:4} = (\boldsymbol{S}^{(0)}_5, \boldsymbol{S}^{*(1)}_{2,0:4}, \boldsymbol{S}^{*(2)}_{2,1:4})^T$ and $\boldsymbol{S}^*_{3,0;3,1:4} = (\boldsymbol{S}^{(0)}_5, \boldsymbol{S}^{*(1)}_{3,0:4}, \boldsymbol{S}^{*(2)}_{3,1:4})^T$ with

$$
\begin{aligned}
\boldsymbol{S}^{*(1)}_{2,0:4} &= \frac{3\times 5}{25}\boldsymbol{S}^{(1)}_{2,0:5} + \frac{3\times 0}{25}\boldsymbol{S}^{(1)}_{2,1:5} + \frac{2\times 5}{25}\boldsymbol{S}^{(1)}_{3,0:5} + \frac{2\times 0}{25}\boldsymbol{S}^{(1)}_{3,1:5}, \\
\boldsymbol{S}^{*(2)}_{2,1:4} &= \frac{3\times 4}{25}\boldsymbol{S}^{(2)}_{2,1:5} + \frac{3\times 1}{25}\boldsymbol{S}^{(2)}_{2,2:5} + \frac{2\times 4}{25}\boldsymbol{S}^{(2)}_{3,1:5} + \frac{2\times 1}{25}\boldsymbol{S}^{(2)}_{3,2:5}, \\
\boldsymbol{S}^{*(1)}_{3,0:4} &= \frac{2\times 5}{25}\boldsymbol{S}^{(1)}_{3,0:5} + \frac{2\times 0}{25}\boldsymbol{S}^{(1)}_{3,1:5} + \frac{3\times 5}{25}\boldsymbol{S}^{(1)}_{4,0:5} + \frac{3\times 0}{25}\boldsymbol{S}^{(1)}_{4,1:5}, \\
\boldsymbol{S}^{*(2)}_{3,1:4} &= \frac{2\times 4}{25}\boldsymbol{S}^{(2)}_{3,1:5} + \frac{2\times 1}{25}\boldsymbol{S}^{(2)}_{3,2:5} + \frac{3\times 4}{25}\boldsymbol{S}^{(2)}_{4,1:5} + \frac{3\times 1}{25}\boldsymbol{S}^{(2)}_{4,2:5}.
\end{aligned}
$$

Then, we clearly have

$$\boldsymbol{S}^* = \frac{1}{50}\begin{pmatrix} 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 & 50 \\ 0 & 0 & 15 & 35 & 50 & 50 & 0 & 15 & 35 & 50 & 50 & 15 & 35 & 50 & 50 & 35 & 50 & 50 & 50 & 50 & 50 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12 & 28 & 40 & 40 & 15 & 35 & 50 & 50 & 35 & 50 & 50 & 50 & 50 & 50 \end{pmatrix}^T.$$

Similarly, the multi-state survival signature of its equivalent system of size 6 can be given from Theorem 5.6 as

$$\boldsymbol{S}^* = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & \frac{1}{5} & \frac{1}{2} & \frac{4}{5} & 1 & 1 & 0 & \frac{1}{5} & \frac{1}{2} & \frac{4}{5} & 1 & 1 & \frac{1}{5} & \frac{1}{2} & \frac{4}{5} & 1 & 1 & \frac{1}{2} & \frac{4}{5} & 1 & 1 & \frac{4}{5} & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{15} & \frac{1}{3} & \frac{8}{15} & \frac{2}{3} & \frac{2}{3} & \frac{14}{75} & \frac{7}{15} & \frac{56}{75} & \frac{14}{15} & \frac{14}{15} & \frac{1}{2} & \frac{4}{5} & 1 & 1 & \frac{4}{5} & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T.$$

The same result can also be obtained by using Theorem 5.3 twice; see Sect. 3 in Yi et al. [71] for details.

8 Concluding Remarks

Signature theory plays an important role in the field of reliability. Research based on signature theory has been reviewed in this work focusing especially on signature related concepts, properties of these concepts, their computational methods, and some multi-state signatures for multi-state systems. On the topic of multi-state signature for multi-state systems, the presented discussions can be summarized as follows: (1) concepts such as multi-state signature, multi-state joint signature, dynamic multi-state signature, ordered multi-state survival signature have been introduced and their properties have been discussed; (2) Stochastic comparisons have been presented for multi-state systems of different sizes based on multi-state signature, multi-state joint signature and multi-state survival signature; and (3) Multi-state systems based on a structure of modules have also been considered based on multi-state signatures and multi-state survival signatures of modules and structures. It will be of great interest to consider statistical inferential problems based on these multi-state signture concepts, especially under nonhomogeneous Poisson processes as in [18, 74]. We are currently working in this direction and hope to report the findings in a future paper.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 72001016), the Fundamental Research Funds for the Central Universities (buctrc202102) and the Funds for First-class Discipline Construction (XK1802-5), and the Natural Sciences and Engineering Research Council of Canada (to the second author) through an Individual Discovery Grant (RGPIN-2020-06733). Our special thanks also go to the editors of this volume for extending an invitation to present this article for the volume.

References

1. Amini-Seresht E, Khaledi BE, Kochar S (2020) Some new results on stochastic comparisons of coherent systems using signatures. J Appl Probab 57(1):156–173

2. Ashrafi S, Asadi M (2014) Dynamic reliability modeling of three-state networks. J Appl Probab 51(4):999–1020
3. Ashrafi S, Asadi M (2015) On the stochastic and dependence properties of the three-state systems. Metrika 78(3):261–281
4. Ashrafi S, Asadi M, Navarro J (2022) Joint reliability function of coherent systems with shared heterogeneous components. Methodol Comput Appl Probab 24:1485–2022
5. Balakrishnan N, Volterman W (2014) On the signatures of ordered system lifetimes. J Appl Probab 51(1):82–91
6. Bedford T, Quigley J, Walls L, Alkali B, Daneshkhah A, Hardman G (2008) Advances in mathematical modeling for reliability. IOS Press, Amsterdam
7. Block H, Dugas M R, Samaniego F J. Characterizations of the relative behavior of two systems via properties of their signature vectors. In: Advances in distribution theory, order statistics, and inference (Chapter), pp 279–289 (2006)
8. Boland PJ (2001) Signatures of indirect majority systems. J Appl Probab 38(2):597–603
9. Boland PJ, Samaniego FJ (2004) The signature of a coherent system and its applications in reliability. In: Mathematical reliability: an expository perspective (Chapter), pp 3–30
10. Cerqueti R, Spizzichino F (2015) Signatures of systems with non-exchangeable lifetimes: some implications in the analysis of financial risk. In: Theory and practice of risk assessment (Chapter), pp 361–375
11. Coolen FPA, Coolen-Maturi T (2012) Generalizing the signature to systems with multiple types of components. In: Complex systems and dependability (Chapter), pp 115–130
12. Coolen-Maturi T, Coolen FPA, Balakrishnan N (2021) The joint survival signature of coherent systems with shared components. Reliab Eng Syst Saf 207:107350
13. Da GF, Chan PS, Xu MC (2018) On the signature of complex system: a decomposed approach. Euro J Oper Res 265(3):1115–1123
14. Da GF, Hu TZ (2013) On bivariate signatures for systems with independent modules. In: Stochastic orders in reliability and risk (Chapter), pp 143–166
15. Da GF, Xia LY, Hu TZ (2014) On computing signatures of k-out-of-n systems consisting of modules. Method Comput Appl Probab 16(1):223–233
16. Da GF, Xu MC, Chan PS (2018) An efficient algorithm for computing the signatures of systems with exchangeable components and applications. IISE Trans 50(7):584–595
17. De Costa Bueno V (2013) A multistate monotone system signature. Stat Probab Lett 83(11):2583–2591
18. De Costa Bueno V, Balakrishnan N (2022) A cumulative residual inaccuracy measure for coherent systems at component level and under nonhomogenous Poisson processes. Probab Eng Inf Sci 36(2):294–319
19. Ding WY, Fang R, Zhao P (2021) An approach to comparing coherent systems with ordered components by using survival signatures. IEEE Trans Reliab 70(2):495–506
20. Elperin T, Gertsbakh I, Lomonosov M (1991) Estimation of network reliability using graph evolution models. IEEE Trans Reliab 40(5):572–581
21. Eryilmaz S, Tuncel A (2015) Computing the signature of a generalized k-out-of-n system. IEEE Trans Reliab 64(2):766–771
22. Eryilmaz S, Tuncel A (2016) Generalizing the survival signature to unrepairable homogeneous multi-state systems. Naval Res Logistics 63(8):593–599
23. Franko C, Yalcin F (2017) Signatures of series and parallel systems consisting of non disjoint modules. Commun Stat-Theory Methods 46(23):11425–11439
24. Gertsbakh IB, Shpungin Y (2010) Models of network reliability: analysis, combinatorics, and Monte Carlo. CRC Press, New York
25. Gertsbakh I, Shpungin Y (2012) Stochastic models of network survivability. Qual Technol Quant Manage 9(1):45–58
26. Gertsbakh I, Shpungin Y, Spizzichino F (2011) Signatures of coherent systems built with separate modules. J Appl Probab 48(3):843–855
27. Gertsbakh I, Shpungin Y, Spizzichino F (2012) Two-dimensional signatures. J Appl Probab 49(2):416–429

28. Jia XJ, Shen JY, Xu FQ, Ma RH, Song XY (2019) Modular decomposition signature for systems with sequential failure effect. Reliab Eng Syst Saf 189:435–444
29. Kochar S, Mukerjee H, Samaniego FJ (1999) The signature of a coherent system and its application to comparisons among systems. Naval Res Logistics 46(5):507–523
30. Koutras MV, Triantafyllou IS, Eryilmaz S (2016) Stochastic comparisons between lifetimes of reliability systems with exchangeable components. Methodol Comput Appl Probab 18(4):1081–1095
31. Kumar A, Singh SB (2018) Signature reliability of linear multi-state sliding window system. Int J Qual Reliab Manage 35(10):2403–2413
32. Levitin G, Gertsbakh I, Shpungin Y (2011) Evaluating the damage associated with intentional network disintegration. Reliab Eng Syst Saf 96(4):433–439
33. Lindqvist BH, Samaniego FJ, Huseby AB (2016) On the equivalence of systems of different sizes, with applications to system comparisons. Adv Appl Probab 48(2):332–348
34. Lisnianski A, Frenkel I (2012) Recent advances in system reliability: signatures, multi-state systems and statistical inference. Springer, London
35. Liu YM, Shi YM, Bai XC, Liu B (2018) Stress-strength reliability analysis of multi-state system based on generalized survival signature. J Comput Appl Math 342:274–291
36. Mahmoudi M, Asadi M (2011) The dynamic signature of coherent systems. IEEE Trans Reliab 60(4):817–822
37. Marichal JL (2014) Subsignatures of systems. J Multivar Anal 124:226–236
38. Marichal JL (2015) Algorithms and formulas for conversion between system signatures and reliability functions. J Appl Probab 52(2):490–507
39. Marichal JL, Mathonet P, Navarro J, Paroissin C (2017) Joint signature of two or more systems with applications to multistate systems made up of two-state components. Euro J Oper Res 263(2):559–570
40. Marichal JL, Mathonet P, Spizzichino F (2015) On modular decompositions of system signatures. J Multivar Anal 134:19–32
41. Marichal JL, Mathonet P, Tamás W (2011) On signature-based expressions of system reliability. J Multivar Anal 102(10):1410–1416
42. Mohammadi L (2017) The joint signature of parallel systems for different permutations of failure times. Comput Stat 32(4):1727–1746
43. Mohammadi L (2017) The joint signature of coherent systems. Naval Res Logistics 64(7):566–579
44. Mohammadi L (2019) The joint reliability signature of order statistics. Commun Stat-Theory Methods 48(19):4730–4747
45. Naqvi S, Chan PS, Mishra DB (2022) System signatures: a review and bibliometric analysis. Commun Stat-Theory Methods 51(7):1993–2008
46. Navarro J, Balakrishnan N, Samaniego FJ (2008) Mixture representations of residual lifetimes of used systems. J Appl Probab 45(4):1097–1112
47. Navarro J, Fernandez-Sanchez J (2020) On the extension of signature-based representations for coherent systems with dependent non-exchangeable components. J Appl Probab 57:429–440
48. Navarro J, Rubio R (2010) Computations of signatures of coherent systems with five components. Commun Stat-Simul Comput 39(1):68–84
49. Navarro J, Ruiz JM, Sandoval CJ (2005) A note on comparisons among coherent systems with dependent components using signatures. Stat Probab Lett 72(2):179–185
50. Navarro J, Ruiz JM, Sandoval CJ (2007) Properties of coherent systems with dependent components. Commun Stat-Theory Methods 36(1):175–191
51. Navarro J, Rychlik T (2007) Reliability and expectation bounds for coherent systems with exchangeable components. J Multivar Anal 98(1):102–113
52. Navarro J, Samaniego FJ, Balakrishnan N (2010) The joint signature of coherent systems with shared components. J Appl Probab 47(1):235–253
53. Navarro J, Samaniego FJ, Balakrishnan N (2011) Signature-based representations for the reliability of systems with heterogeneous components. J Appl Probab 48(3):856–867

54. Navarro J, Samaniego FJ, Balakrishnan N (2013) Mixture representations for the joint distribution of lifetimes of two coherent systems with shared components. Adv Appl Probab 45(4):1011–1027
55. Navarro J, Samaniego FJ, Balakrishnan N, Bhattacharya D (2008) On the application and extension of system signatures in engineering reliability. Naval Res Logistics 55(4):313–327
56. Navarro J, Spizzichino F, Balakrishnan N (2010) Applications of average and projected systems to the study of coherent systems. J Multivar Anal 101(6):1471–1482
57. Reed S (2017) An efficient algorithm for exact computation of system and survival signatures using binary decision diagrams. Reliab Eng Syst Saf 165:257–267
58. Reed S, Lofstrand M, Andrews J (2019) An efficient algorithm for computing exact system and survival signatures of K-terminal network reliability. Reliab Eng Syst Saf 185:429–439
59. Samaniego FJ (1985) On closure of the IFR class under formation of coherent systems. IEEE Trans Reliab 34(1):69–72
60. Samaniego FJ (2007) System signatures and their applications in engineering reliability. Springer, New York
61. Samaniego FJ, Balakrishnan N, Navarro J (2009) Dynamic signatures and their use in comparing the reliability of new and used systems. Naval Res Logistics 56(6):577–591
62. Samaniego FJ, Navarro J (2016) On comparing coherent systems with heterogeneous components. Adv Appl Probab 48(1):88–111
63. Triantafyllou IS, Koutras MV (2008) On the signature of coherent systems and applications. Probab Eng Inf Sci 22(1):19–35
64. Yi H, Cui LR (2018) A new computation method for signature: Markov process method. Naval Res Logistics 65(5):410–426
65. Yi H, Balakrishnan N, Cui LR (2020) On the multi-state signatures of ordered system lifetimes. Adv Appl Probab 52(1):291–318
66. Yi H, Balakrishnan N, Cui LR (2021) Comparisons of multi-state systems with binary components of different sizes. Methodol Comput Appl Probab 23(4):1309–1321
67. Yi H, Balakrishnan N, Cui LR (2022) On dependent multi-state semi-coherent systems based on multi-state joint signature. Methodol Comput Appl Probab 24(3):1717–1734
68. Yi H, Balakrishnan N, Li X (2023) On ordered system signature and its dynamic version for coherent systems with applications. J Appl Probab 60(3):AP-18485 (accepted)
69. Yi H, Balakrishnan N, Li X (2022) Signatures of multi-state systems based on a series/parallel/recurrent structure of modules. Probab Eng Inf Sci 36(3):824–850
70. Yi H, Balakrishnan N, Li X (2022) Equivalency in joint signatures for binary/multi-state systems of different sizes. Probab Eng Inf Sci 36(4):1027–1054
71. Yi H, Balakrishnan N, Li X (2022) Equivalency of multi-state survival signatures of multi-state systems of different sizes and its use in the comparison of systems. Probab Eng Inf Sci. https://doi.org/10.1017/S0269964822000183
72. Yi H, Balakrishnan N, Li X (2022) Ordered multi-state system signature and its dynamic version in evaluating used multi-state systems. Probab Eng Inf Sci. https://doi.org/10.1017/S0269964822000237
73. Yi H, Cui LR, Balakrishnan N (2021) Computation of survival signatures for multi-state consecutive-k systems. Reliab Eng Syst Saf 208:107429
74. Zarezadeh S, Asadi M, Balakrishnan N (2014) Dynamic network reliability modeling under nonhomogeneous Poisson processes. Euro J Oper Res 232(3):561–571
75. Zarezadeh S, Asadi M, Eftekhar S (2019) Signature-based information measures of multi-state networks. Probab Eng Inf Sci 33(3):438–459
76. Zarezadeh S, Mohanmmadi L, Balakrishnan N (2018) On the joint signature of several coherent systems with some shared components. Euro J Oper Res 264(3):1092–1100

Comprehensive Reliability of Aircraft Actuation System

Shaoping Wang, Jian Shi, and Yajing Qiao

Abstract Aircraft actuation system receives commands from the flight control computer and drives the plane surface to realize the aircraft flight attitude and flight trajectory control. The actuation system has a significant influence on the overall aircraft flight control performance and safety. This chapter presents the essential reliability characteristics of the redundancy aircraft actuation systems, and creates a reliability evaluation method for non-similar redundancy actuation systems. Section 1 summarizes the aircraft actuation system, and explains the interface between the flight control system(FCS) and the actuation system. Some typical aircraft hydraulic actuation system constructions in current commercial aircraft are provided in Sect. 2. Furthermore, Sect. 3 analyzes the architecture and characteristics of A380 aircraft actuation system, and provides the comprehensive reliability definition and the reliability calculation method. Afterward, the reliability of actuation system based on performance degradation is described. Finally, the integrated reliability evaluation case is provided for example of a HA/EHA system, in which HA operates actively and EHA follows under normal operating conditions.

Keyword Reliability · Aircraft actuation system · Flight control system · Performance degradation · HA/EHA system

1 Main Properties of Aircraft Actuation Systems

There are three key elements for an aircraft, which are lift force, power and manipulation. Aircraft aerodynamic configuration determines aircraft available lift force. The engine provides aircraft thrust and onboard power supply, and the flight control system (FCS) provides aircraft translational and rotational motion by driving the control surfaces [1]. The control surfaces are used directly to control the attitude of the aircraft. The control surfaces on aircraft, illustrated in Fig. 1, are generally divided into two groups in terms of their contribution to aircraft flight control, i.e., primary

S. Wang (✉) · J. Shi · Y. Qiao
School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
e-mail: shaopingwang@vip.sina.com

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_9

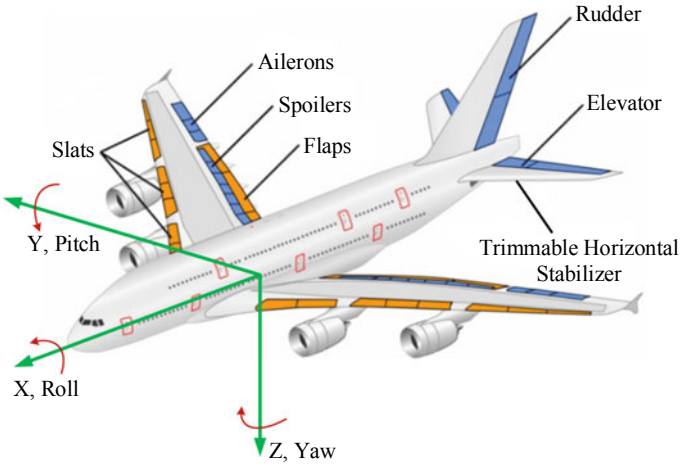


Fig. 1 Control surfaces of an advanced commercial aircraft

control surfaces and secondary control surfaces. The primary flight control surfaces include elevators (for pitch control), rudders (for yaw control) and ailerons (for roll control), which are dedicated to the control of the roll, yaw and pitch attitudes and of the trajectory of the aircraft [2]. The secondary flight control surfaces consist of slates, spoilers, flaps et al., which are dedicated to the control of the lift of the wing, and help aircraft control during take-off or landing [3].

All of the control surfaces on aircraft are driven by several actuators which are powered by onboard hydraulic or electric supply. Figure 2 shows a typical schematic of aircraft flight control and actuation system [4], which consists of flight control computer (FCC), actuator control electronics (ACE), actuator and sensors. Under normal operation, FCCs receive pilot control commands, and determine the control surface movement for the aircraft to respond in the best way to the pilot's commands and achieve a fast, well damped response throughout the flight envelope. Then, FCCs transmit actuation commands to ACE and actuators to drive the control surface to desired position. Meanwhile, sensors, e.g. inertial measurement unit (IMU) and air data sensors, measure the aircraft response and close the pitch, roll, and yaw control loops to ensure that the aircraft possesses well-harmonized control characteristics throughout the flight regime [5].

The architecture of the flight control and actuation system, in terms of number of actuators per surface, number and distribution of power sources and flight control computers, is primarily driven by safety considerations. The safety objectives, as defined by the current regulations, require failures, or combinations of failures, resulting in the loss of the aircraft to be demonstrated as "Extremely Improbable". This means that their failure rate shall not exceed a probability of 10^{-9} per flight hour. Complete loss of power supply to a fully powered flight control and actuation system, which would result in loss of control, falls in this category. As a consequence, the

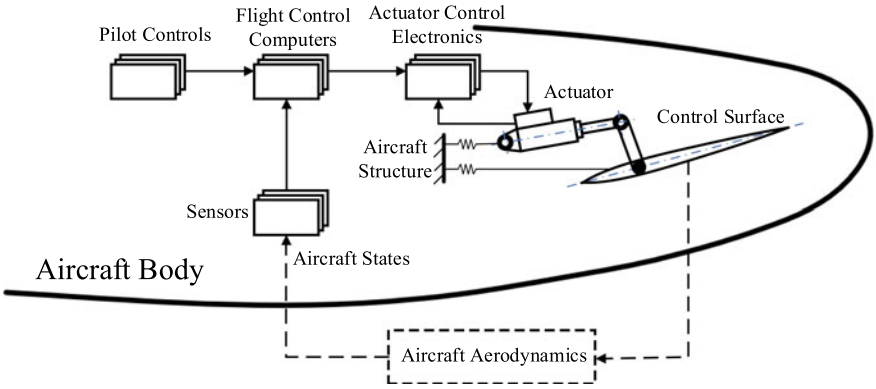


Fig. 2 Aircraft flight control system

flight control and actuation system shall be supplied from several redundant power sources [6].

2 Introduction of Aircraft Actuators

Actuation system is a vital link between the flight control and power supply, providing the motive force necessary to move flight control surfaces. Since the flight controls need the force to drive the surface motion, several sets of hydraulic supply and electrical supply system are used in aircraft. With the development of more electric aircraft, diverse actuators, including hydraulic actuator (HA), electro-hydrostatic actuator (EHA), Electro-mechanical actuator (EMA), and electrical backup hydraulic actuators (EBHA), are used in in aircraft. Among these types of actuators, application of EMA is limited due to its airworthiness uncertainty.

2.1 Hydraulic Actuator (HA)

HA, also called valve-controlled hydraulic actuator, is widely used in aircraft. It is a device converting hydraulic power to mechanical power through the control element [7]. Within Hydraulic actuator, a servo-valve (SV) serves as a power interface between electrical and hydraulic powers. The basic schematic of hydraulic actuator is illustrated in Fig. 3.

The digital fly-by-wire (FBW) commands or direct electrical link demands from the flight deck controls are processed by the ACE which supplies an analogue command to the actuator SV. This allows aircraft hydraulic power to be supplied

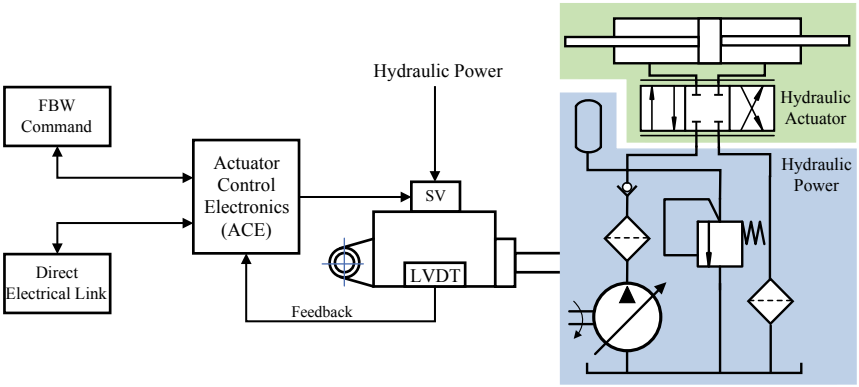


Fig. 3 Hydraulic actuation system

to the appropriate side of cylinder ram piston moving the ram to the desired position. In this implementation, the cylinder ram position is detected by means of a Linear Variable Differential Transducer (LVDT) which feeds the signal back to the ACE where the loop around the actuator is closed. The conventional HA is continually pressurized by a centralized hydraulic power supply system whether or not there is any demand whereas the actuator demands are minimal in many cases. In this situation, most of the energy coming from the hydraulic power supply system is converted to heat through the orifice. Therefore, the constant pressure supply actuator is efficient in terms of dynamic response but not efficient in terms of energy.

Modern flight controls are increasingly adopting FBW system, in which the electrical parts replace the mechanical one. As vital elements in a FBW system, actuation system must be able to survive any two failures and carry on operating satisfactorily in order to meet the aircraft safety requirements. Since the limited reliability of electrical control part of SV, aircraft utilizes multiple redundancy actuation system. Figure 4 illustrates a simplified block schematic diagram of multiple redundant hydraulic actuator, in which four identical lanes solenoid (SOL) are used. The SVs summed before applying demands to the control valves. This kind of multiple redundant actuator is also called quadruplex actuation system [5].

2.2 Electro-Hydrostatic Actuator (EHA) [8]

The Electro-Hydrostatic Actuator (EHA) uses state-of-the-art power electronics and control techniques to provide more efficient flight control actuation. EHA transmits power to the load through a hydrostatic loop which involves variable speed motor, fixed displacement hydraulic pump, cylinder and LVDT, as illustrated in Fig. 5. An electric power source directly supplies an AC motor that drives a variable displacement pump connected to a hydraulic cylinder. This makes a hydrostatic loop. The

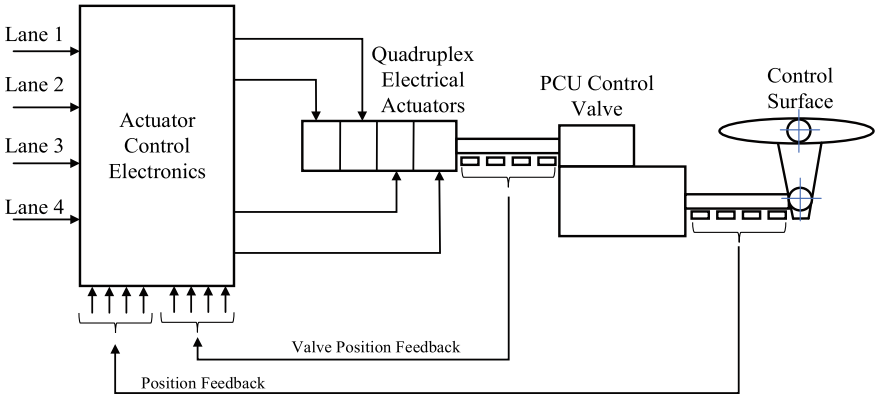


Fig. 4 Quadruplex actuation system

load is closed loop position-controlled by action on the pump displacement. The EHA seeks to provide a more efficient form of actuation where the actuator only draws significant power when a control demand is sought. The EHA accomplishes this by using the three-phase AC power to feed power drive electronics which in turn drive a variable speed pump together with a constant displacement hydraulic pump. When a demand is received from the ACE, the power drive electronics is able to react sufficiently rapidly to drive the variable speed motor and hence pressurize the actuator such that the associated control surface may be moved to satisfy the demand. Once the demand has been satisfied then the power electronics resumes its normal dormant state. Consequently, power is only drawn from the aircraft buses bars while the actuator is moving, representing a great saving in energy. The ACE closes the control loop around the actuator electrically as previously described.

The EHA uses the local hydraulic system, which reduces the need for long pipes between the centralized hydraulic power supply and the actuator, and thus decreases the corresponding weight [9]. In addition, in case of no demand, the only power requirement of EHA is to maintain the control electronics. When the actuator control

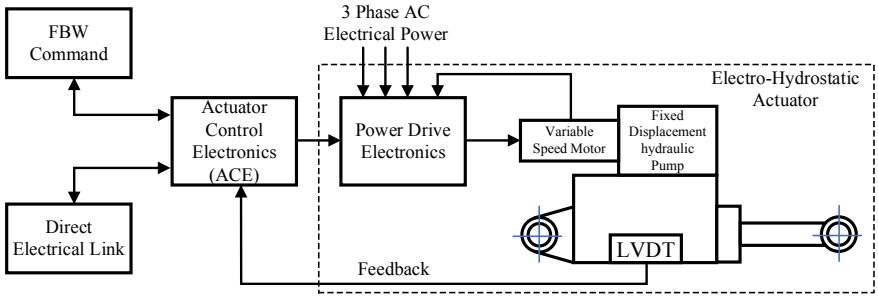


Fig. 5 Electro-hydrostatic actuator (EHA)

equipment sends the command, the power rapidly acts on the electronics to drive the variable speed motor and pressurizes the actuator resulting in the corresponding surface movement. Once the output of the surface satisfies the demand, the power electronic resumes its normal dormant state. The ACE electrically closes the control loop around the actuator. It is obvious that the power of EHA is only drawn from the aircraft buses while the actuator is moving, so EHA can save energy.

2.3 HA/EHA System

In order to improve the reliability, the heterogeneous dissimilar redundant actuation system, namely HA/EHA system [8], is used to drive one piece of surface simultaneously in more electrical aircraft shown in Fig. 6.

In Fig. 6, HA accomplished with EHA constitutes the dissimilar redundant actuator system to drive the control surface together. To a HA/EHA system, there are two operating modes: HA active/EHA standby and HA fault/EHA active. In normal condition, HA drives control surface and EHA follow-up. When hydraulic power supply is lost, or HA fails, EHA substitutes HA to drive control surface [10].

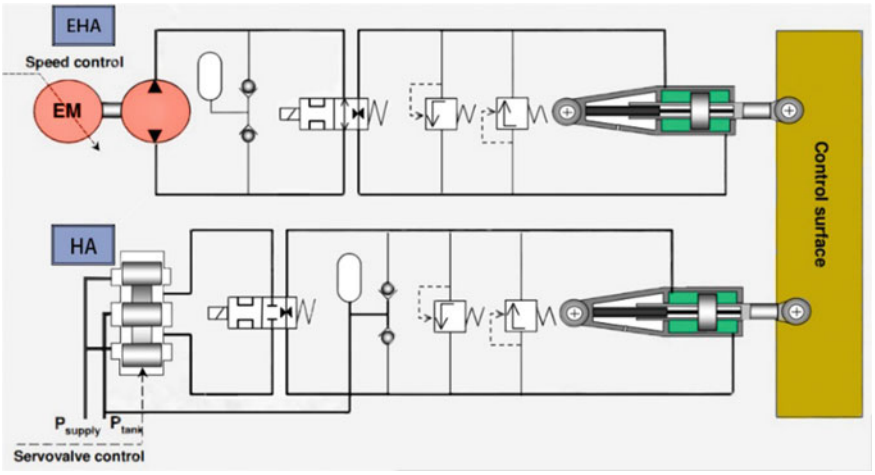


Fig. 6 HA/EHA heterogeneous actuation system

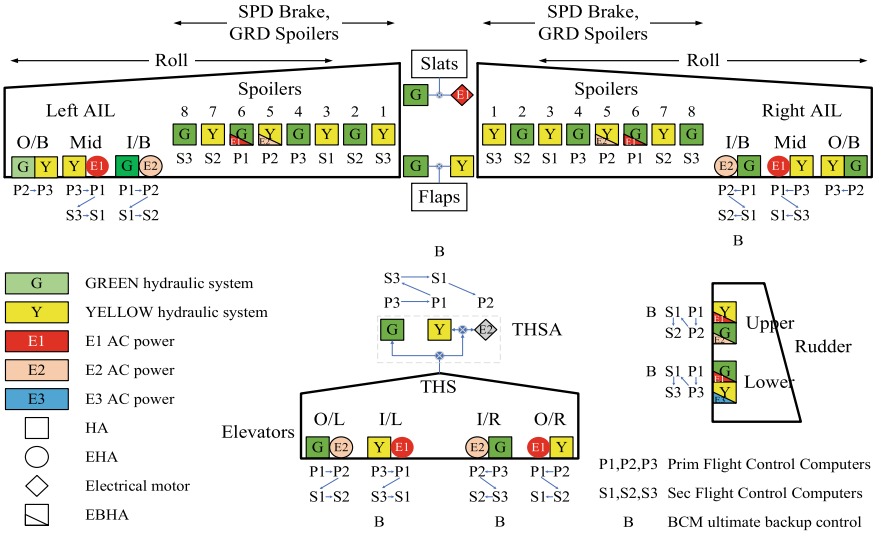
3 Reliability of A380 Aircraft Actuation Systems

3.1 A380 Aircraft Characteristics

Conventional commercial aircraft, including previous Airbus products, are provided with 3 hydraulic systems, while they are also equipped with 2 main electrical systems for supplying other users, which makes a total of 5 power sources on board.

The A380 is a plane equipped with high reliability actuation system based on two sets of hydraulic power supply and two sets of electrical power supply [11]. This type of power source distribution, identified as "2H/2E", features two hydraulic systems, so called Green (G) and Yellow (Y), and two electric systems, E1 and E2, as shown in Fig. 7. The hydraulic actuators are normally active while the electrically powered actuators are normally stand-by and become operative in the event of a failure of the normal, hydraulically supplied, control lane.

As shown in Fig. 7, the outboard aileron surface and spoiler surfaces on each wing are driven by HAs which get power from hydraulic system. The middle, inboard aileron surfaces and the elevator surfaces are driven by HAs and EHAs. In this case, HA is normally active while EHA is normally stand-by and becomes operative in the event of HA failure. The two spoiler surfaces of each wing and rudder are driven by the electrical backup hydraulic actuators (EBHA) with features of HA and EHA. The trimmable horizontal stabilizer actuator is powered independently by hydraulic and electrical power supply.



O-outboard, I- inboard, MIL-middle, L- left, R- right, AIL- aileron, THS- trimmable horizontal stabilizer

Fig. 7 "2H/2E" structure of A380

The driven modes of A380 actuation systems include:

- Normal mode—HA mode: HA receives the commands from ACE, and the actuator SV control hydraulic power to be supplied to the actuator moving the cylinder ram to the desired position.
- Backup mode—EHA mode: EHA receives the commands from ACE, and the electrical motor operates with the electrical power from the aircraft AC electrical system. The electrical motor drives the fixed displacement hydraulic pump to move the actuator cylinder ram to the desired position.

By eliminating one hydraulic system and replacing it with a set of electrical power supply, the A380 philosophy is more toward the electric flight control actuation concept. This approach does have positive effect on the probability of losing the flight control actuation system. With two hydraulic systems and two electric systems, the A380 is more reliable and safer than other types of aircraft designs.

3.2 Reliability Assessment Based on Probability Theory

MIL-HDBK-781 [12] is a standard developed by the US military over a period of years to use an analytical bottom-up approach to predicting reliability. It uses type of component, environment and quality factor as major discriminators in predicting the failure rate of a particular component, module, and ultimately subsystem. Based on the guidelines of MIL-HDBK-781, the reliability assessment architecture for aircraft actuator system is shown in Fig. 8 [13], which is divided into three levels: component reliability model level, subsystem (HA, EHA) reliability model level, and actuation system reliability model level.

Reliability data are an essential part of a probabilistic reliability assessment. The use of generic component reliability data is therefore unavoidable. A relatively large amount of component reliability data is available in the open literature. At the level of component reliability model, the basic actuator component reliability data, such as failure rate, MTBF, and failure probability, is collected from various reliability data handbooks and standard manuals, for example, IAEA-TECDOC-478.

By establishing HA/EHA subsystem reliability model, the reliability of HA/EHA is evaluated at the level of subsystem reliability. The reliability of the entire aircraft actuation system is finally obtained by considering the operating mode and reconfiguration strategies of hydraulic/electric dissimilarity actuation system [14].

Classic probability theory provides a means to analyze the failure probability of this complex actuation systems architectures.

3.2.1 Reliability of Components

Reliability can be expressed either as the probability that an item or system will operate in a satisfactory manner for a specified period of time, or, when used under

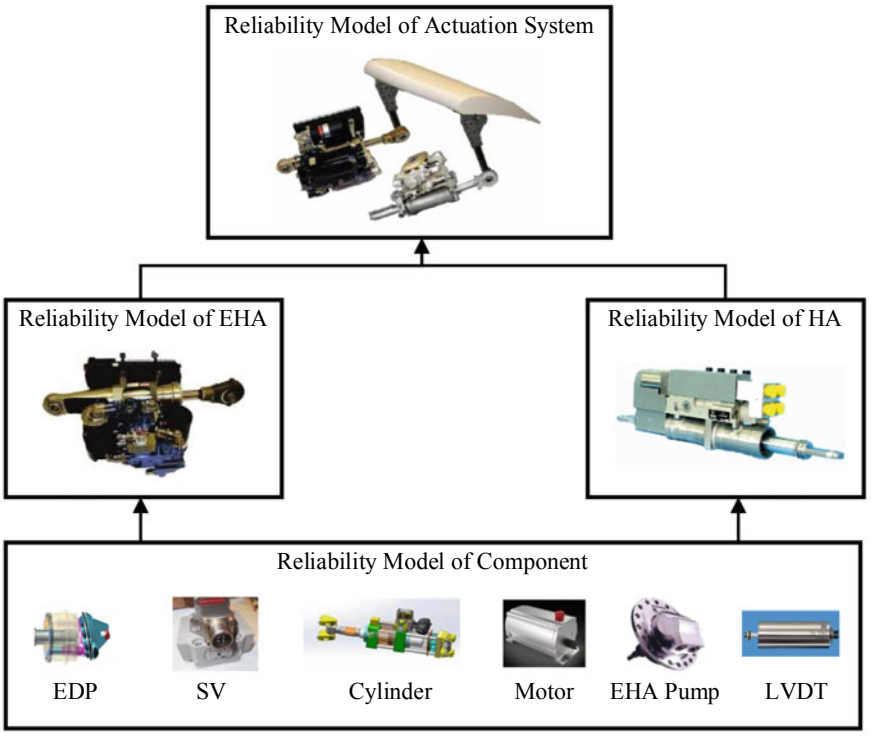


Fig. 8 Aircraft actuation system composition diagram

stated conditions, in terms of its Mean Time between Failures (MTBF) [15]. For repairable systems, such as aircraft actuation system, that must operate continuously, reliability is usually expressed as the probability that a system will perform a required function under specific conditions for a stated period of time.

Suppose the life distribution of component in aircraft actuation system obeys exponential distribution, whose failure rate is shown in Table 1. According to the reliability theory based on probability, the reliability of component under exponential distribution can be described as

$$R(t) = e^{-\lambda t} \tag{1}$$

where λ is the failure rate of component, t is the operational time. Considering the limited flight time of aircraft, the reliability of components can be calculated under $t = 20h$ in Table 1.

Table 1 Failure rate of component in aircraft actuation system

No	Component	Failure rate		Reliability under $t = 20h$	
		Param	Value ($\times 10^{-4}/h$)	Param	Value
1	Engine drive pump (EDP)	λ_{EDP}	1.4521	R_{EDP}	0.9971
2	Electrical motor (EM)	λ_{EM}	1.7029	R_{EM}	0.9966
3	Hydraulic servo valve (SV)	λ_{SV}	1.0603	R_{SV}	0.9979
4	Cylinder (Cy)	λ_{Cy}	0.4502	R_{Cy}	0.9991
5	LVDT	λ_{LVDT}	0.7506	R_{LVDT}	0.9985
6	EHA pump (EP)	λ_{EP}	1.8530	R_{EP}	0.9963

3.2.2 Reliability of Aircraft Actuation System

(1) **Reliability of HA**

To HA, the total lane function will be lost, or no longer available, if any of the component fails. The reliability of HA is described as

$$R_{HA}(t = 20h) = R_{EDP} \cdot R_{SV} \cdot R_{Cy} \cdot R_{LVDT} = 0.9926 \tag{2}$$

The failure rate of HA is expressed as

$$\lambda_{HA} = \lambda_{EDP} + \lambda_{SV} + \lambda_{Cy} + \lambda_{LVDT} = 3.71 \times 10^{-4}/h \tag{3}$$

(2) **Reliability of EHA**

HA consists of electrical motor, high speed pump, cylinder and LVDT, whose reliability is calculated as

$$R_{EHA}(t = 20h) = R_{EM} \cdot R_{EP} \cdot R_{Cy} \cdot R_{LVDT} = 0.9905 \tag{4}$$

The failure rate of EHA is expressed as

$$\lambda_{EHA} = \lambda_{EM} + \lambda_{EP} + \lambda_{Cy} + \lambda_{LVDT} = 4.76 \times 10^{-4}/h \tag{5}$$

(3) **Reliability of HA/HA**

To a redundant actuation system with two HAs in which one HA is powered by green power supply and another is powered by yellow power supply, its reliability is described as

$$R_{HA/HA} = 1 - \left[1 - e^{-\lambda_{HA}t}\right]^2 = 0.99994524 \tag{6}$$

The failure rate of HA/HA is shown as

$$\lambda_{HA/HA} = \frac{2\lambda_{HA}e^{-\lambda_{HA}t} - 2\lambda_{HA}e^{-2\lambda_{HA}t}}{2e^{-\lambda_{HA}t} - e^{-2\lambda_{HA}t}} = 5.45 \times 10^{-6}/h \tag{7}$$

(4) **Reliability of HA/EHA**

To an aircraft hydraulic/electric dissimilarity actuator system consisting of HA and EHA with the active/standby operation mode, its reliability is described as

$$R_{HA/EHA} = e^{-\lambda_{HA}t} + \frac{\lambda_{HA}}{\lambda_{HA} - \lambda_{EHA}}\left[e^{-\lambda_{EHA}t} - e^{-\lambda_{HA}t}\right] = 0.99993 \tag{8}$$

Its mean time to failure is calculated as

$$MTTF_{HA/EHA} = \frac{1}{\lambda_{HA}} + \frac{1}{\lambda_{EHA}} \tag{9}$$

3.2.3 Reliability of Aircraft Primary Control System

(1) **Mission reliability** [16]

Mission reliability is defined as the probability of the system for being free of failure for the period of time required to complete a mission. The probability is a point on the reliability function corresponding to the mission length. The mission reliability of a system can be described as

$$R_M(t) = P(T > t_M) \tag{10}$$

where $R_M(t)$ is the mission reliability of system, P is the probability, T is the life of system, and t_M is the mission time.

There are two indices to evaluate the reliability of aircraft actuation system: flight safety and mission reliability. According to the aircraft control system design specification (MIL-F-9490D) [17], the probability of mission failure per flight due to relevant material failures in the FCS shall not exceed the applicable limit specified below.

- Overall aircraft mission accomplishment reliability is specified by the procurement activity

$$Q_{M(FCS)} \leq (1 - R_M)A_{M(FCS)} \tag{11}$$

- Overall aircraft mission accomplishment reliability is not specified

$$Q_{M(FCS)} \leq 1 \times 10^{-3} \tag{12}$$

Table 2 FCS Quantitative flight safety requirements

MIL-F-8785, class III aircraft	Maximum aircraft loss rate from FCS failure	
	Param	Value
All rotary wing aircraft	$Q_{S(FCS)}$	5×10^{-7}
MIL-F-8785 class I, II, and IV aircraft	$Q_{S(FCS)}$	5×10^{-7}
MIL-F-8785, class III aircraft	$Q_{S(FCS)}$	100×10^{-7}

where $Q_{M(FCS)}$ is the maximum acceptable mission unreliability due to relevant FCS material failures, R_M is the specified overall aircraft mission accomplishment reliability, and $A_{M(FCS)}$ is the mission accomplishment allocation factor for flight control (chosen by the contractor).

Failures in power supplies or other subsystems that do not otherwise cause aircraft loss shall be considered where pertinent. A representative mission to which the requirement applied should be established and defined in the FCS specification.

(2) **Safety requirement**

If the overall aircraft flight safety in terms of R_S is not specified by the procuring activity, the numerical requirements given in Table 2 [18].

The probability of aircraft loss per flight due to relevant FCS material failures in the FCS shall not exceed $Q_{S(FCS)} \leq (1 - R_S)A_{S(FCS)}$, where $Q_{S(FCS)}$ is the maximum acceptable aircraft loss rate due to relevant FCS material failures, R_S is the specified overall aircraft flight safety requirement as specified by the procuring activity, and $A_{S(FCS)}$ is the flight safety allocation factor for flight control (chosen by the contractor).

At present, the safety requirement of a FCS is 1.0×10^{-7}/flight hour for military aircraft and $1 \times 10^{-9} \sim 1 \times 10^{-10}$/flight hour for commercial aircraft. To achieve such high reliability requirements, it is necessary to utilize the redundancy design method.

The overall reliability of the aircraft hydraulic actuation system depends on the computer control/monitor architecture, which provides the tolerance to hardware and software failures, the servo control, and the power supply arrangement. Thus, the redundancy, failure monitoring, and system protection emerged in the system design. The aircraft safety is demonstrated in the airworthiness regulation. In aircraft design, the faults, interaction faults, and external environmental hazards should be considered.

(3) **Reliability of primary control system**

Primary flight control includes pitch, roll and yaw, in which the corresponding manipulate surface are elevator, aileron and rudder. In Fig. 7, the elevator and rudder adopt dual redundant EHAs respectively. To elevator, two bilateral symmetrical sub-surfaces are driven by 2HA/EHAs. These surfaces provide very powerful pitch control authority by an agile high performance aircraft. Without the benefit of reliable computer driven control system, the aircraft would be uncontrollable and would

Table 3 Reliability of primary flight control system

Flight controls	Surface composition (L/R)	Reliability
Pitch	2HA/EHA	0.999999995
Roll	HA/HA + 2HA/EHA	0.999999998
Yaw	2HA/EHA	0.999999995
Primary control system	Pitch + Roll + Yaw	0.99999943

crash in a matter of seconds. The roll control is provided the differential motion of the aileron, augmented to a degree by spoilers, in which the aileron includes HA/HA + HA/EHA + HA/EHA and spoiler composes of HA/HA + 2HA/EHA + 2HA/HA. To some extent roll controls are also provided by differential action of the wing trailing edge spoilers. However, most of the roll controls are provided by ailerons. Yaw control is provided by the upper and lower rudder, in which the sub-surfaces are driven by 2 EHAs. Using the reliability theory based on probability, the reliability of three-axis control system can be show in Table 3.

From Table 3, the equivalent failure rate of primary flight control system is approximately $10^{-6}/h$, which meets the safety requirement shown in Table 2.

3.3 Comprehensive Reliability of Actuation System

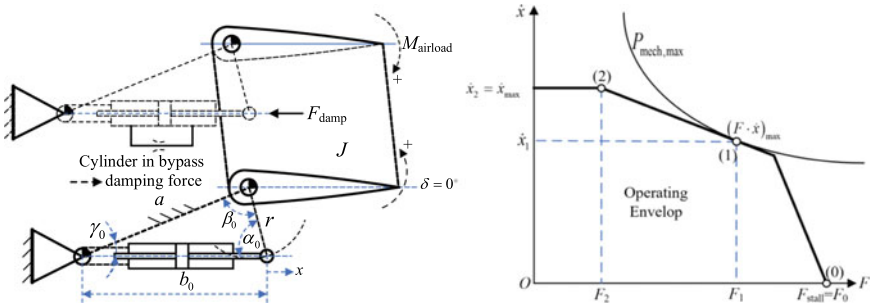
3.3.1 Dynamic Performance of Redundant Actuation System

Since the safety and reliability of flight control system is very important, the redundancy techniques are widely used in actuation system design [8]. Critical surfaces (e.g., aileron, elevator, rudder) are driven by two or three actuators, each powered by an independent hydraulic or electric distribution network. These can be operated in alternating mode (active/standby mode) or in simultaneous mode (active/active mode). Any failure of the neighboring actuators must not lead to a loss of control of the movable surface with the remaining operable actuator. To decrease the influence among actuators of the same surface, some functions such as bypass, movement damping, and load limiting should be implemented. If actuators simultaneously drive the surface, load sensing and compensation should be integrated into the system in order to avoid force fighting.

The effective load/speed operating envelope of the actuator is derived from the load profile of the control surface, taking into account the geometric arrangement shown in Fig. 9.

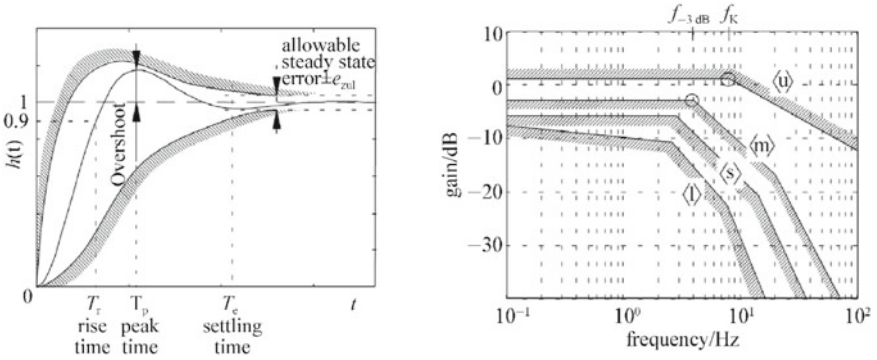
The performance requirements include:

- The stall load $F_{\text{stall}} = F_0$;
- The maximum mechanical power $P_{\text{mech, max}} = \max(F \cdot \dot{x})$ to be provided by the actuator with the load F_1 (include friction) and speed x_1.



(a) Duplex actuator based on HA/EHA (b) actuator operating point

Fig. 9 Redundant actuator structure



(a)Step response of actuator (b) frequency response of actuator

Fig. 10 Dynamic performance requirement of an actuator

- Maximum speed with corresponding load ($\dot{x}_2$, F_2) or no load.

  The dynamic performance requirements are shown in Fig. 10.

3.3.2 Reliability Model Based on Performance Parameter

Aircraft 380 provides 2H/2E power supply and HA/EHA system structure, which has greatly improved the overall performance of the aircraft. However, the reliability assessment of actuation system becomes difficult due to the increased complexity. On the one hand, whether powered by hydraulic or electrical system, the components present a power loss because of performance degradation during the power transmission process. Besides the power loss behaves multi-state with random characteristics affected by operating conditions and other factors. On the other hand, servo actuators need to follow the flight control instructions, which means the power transmission

process is controlled by the flight control computer. It is necessary to describe the reliability of actuation system based on performance degradation.

Suppose the performance degradation parameter as $\boldsymbol{Y}$, the reliability model based on performance degradation can be defined as [16]

$$R_Y(t) = \begin{cases} P\{\mathrm{Y} \geq \Omega | \mathrm{X} \sim \mathrm{D}\}\ \mathrm{Y}\ \text{is tread decreasing} \\ P\{\mathrm{Y} \leq \Omega | \mathrm{X} \sim \mathrm{D}\}\ \mathrm{Y}\ \text{is tread increasing} \end{cases} \tag{13}$$

where Ω is the performance threshold, $\mathrm{X} = \{X_1, X_2, ..., X_n\}$ is the system parameter set, $\boldsymbol{D}$ is the probability distribution of the system parameters.

Actuation system is a typical displacement close-loop system, whose main parameters include rapidity, stability and accuracy. When the component deteriorates with load spectrum, its performance will degrade and lead to failure. Take the rapidity as example, the response time needs to be selected to describe the performance reliability as

$$R_Y(t) = \mathrm{P}\left\{T_\mathrm{r} < T_{r_{threshold}} | \mathrm{X}_j \sim f_{X_j}\left(x_j, t\right)\right\} \tag{14}$$

where $T_{r_{threshold}}$ is the threshold of the response time, X_j is the system parameter set, and $f_{X_j}\left(x_j, t\right)$ is the probability density of system parameters.

3.3.3 Reliability Analysis Based on Performance of Actuation System

Modern aircraft use segmentation control surface technology in flight control systems for high reliability. In some critical divided surfaces, HA/HA and HA/EHA redundant actuation system is adopted to increase the reliability because it uses totally different power supply system and actuator. Figure 11 shows a new kind of application in A380 aircraft that is based on a HA and EHA, in which HA operates actively and EHA follows under normal operating conditions. The EHA assumes operation in the case of HA failure [18].

In Fig. 11, HA section consists of the hydraulic power supply system, SV, cylinder, and LVDT. The input of the SV is i_v, the output force of the HA is F_h, and the displacement of the cylinder is x_h. The EHA section consists of a brushless motor, pump, and cylinder, in which the input to the motor is the control voltage u_e, the output force is F_e, and the displacement of the cylinder is x_e. The inputs to the control surface are the displacement of x_h, x_e, and aerodynamic load F_L, whereas the outputs are the surface displacement x_t and the force acted in both cylinders, F_h and F_e.

(1) **Mathematical model of HA** [19]

In order to get HA mathematical model, the transform function of component is developed as follows.

- **SV model**

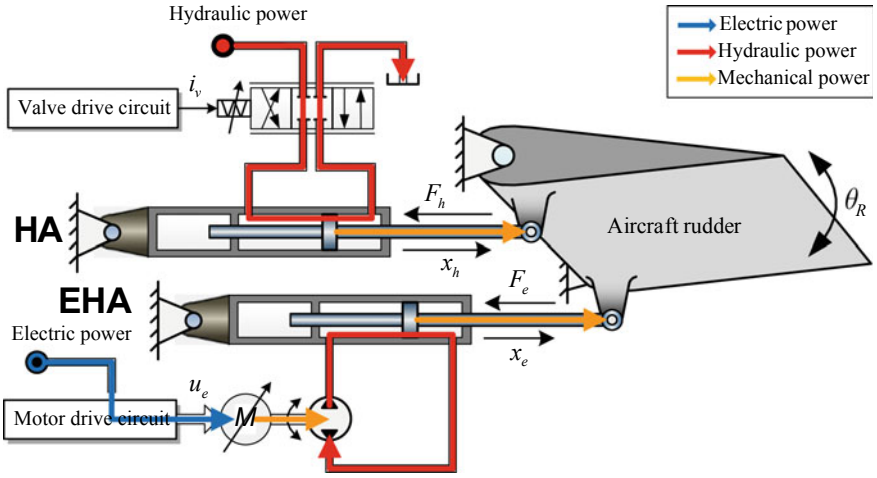


Fig. 11 HA/EHA system

Suppose the SV control current is i_v, the SV spool displacement is x_v, and the gain of the amplifier is K_v. The transfer function of the SV can then be described as a second-order system:

$$Q_h = \frac{K_q K_v \omega_v^2}{s^2 + 2\xi_v \omega_v s + \omega_v^2} i_v - K_c p_h \tag{15}$$

where ω_v is the characteristic frequency of the SV, ξ_v is the damping coefficient of the SV, Q_h is the load flow, p_h is the load pressure, K_q is the flow gain, and K_c is the flow-pressure coefficient.

- **Cylinder model**

Suppose that the fluid has laminar flow, the fluid temperature is constant, and the friction loss and pipe dynamic influence can be neglected. The flow equation of the cylinder can then be written as

$$Q_h = A_h \frac{dx_h}{dt} + \frac{V_{th}}{4E_y} \frac{dp_h}{dt} + C_{sh} p_h \tag{16}$$

The force balance equation of the cylinder is

$$A_h p_h = m_{ph} \frac{d^2 x_h}{dt^2} + B_{ph} \frac{dx_h}{dt} + F_h + f_h \tag{17}$$

where A_h is the piston area, x_h is the displacement of HA, V_{th} is the total volume of HA, E_y is the equivalent volume elastic modulus, C_{sh} is the total leakage coefficient,

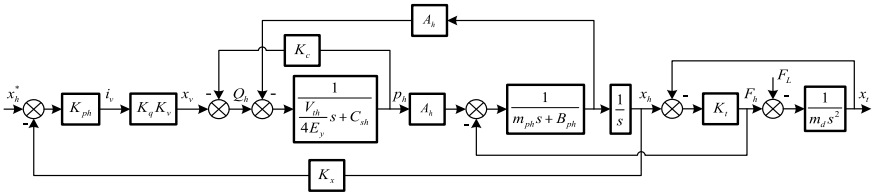


Fig. 12 The block transfer function diagram of HA

m_{ph} is the piston mass, B_{ph} is the viscous damping coefficient, and f_h is the friction between the piston and the cylinder.

- **LVDT model**

Suppose the LVDT coefficient is K_x, then the measured displacement can be shown as

$$x_h' = K_x \cdot x_h \tag{18}$$

The block transfer function diagram of HA is shown in Fig. 12.

(2) **Mathematical model of EHA** [20]

In the EHA channel of the aircraft actuation system, aircraft electrical power drives the brushless DC motor which in turn drives the fixed displacement pump based on local hydraulic system. The detailed mathematical model can be described as follows.

- **DC motor model**

The high-power and high-voltage permanent magnet brushless DC motor for aircraft is the core component of EHA, which provides power for EHA. whose potential balance equation is

$$u = C_e\omega_m + L_e\frac{di_e}{dt} + R_e i_e \tag{19}$$

The torque balance equation of motor can be described as

$$K_m i_e = T_e + J_m\frac{d\omega_m}{dt} + B_m\omega_m \tag{20}$$

where u is control voltage, C_e is back electromotive force coefficient, ω_m is motor speed, L_e is armature inductance, R_e is armature resistance, i is current, K_m is electromagnetic torque constant, T_e is output torque, J_m is the total moment of inertia of motor and pump, B_m is the total load damping coefficient of motor and pump.

- **High speed pump model**

The output torque of DC motor is the input torque of pump T_e:

$$T_e = \frac{D_P}{2\pi}(P_1 - P_2) \tag{21}$$

where D_P is the pump displacement, P_1 and P_2 are pump outlet and inlet pressure.

Set the parameters of variable piston pump as follows: d is piston diameter, D is the diameter of plunger distribution circle, $\alpha_{\max}$ is maximum inclination angle of swashplate, $s_{\max}$ is corresponding maximum stroke of piston, z is piston number. When the output displacement of regulating mechanism is x_P, the displacement of pump can be calculated by [21]

$$D_P = \frac{\pi}{4}d^2(s_{\max} - x_P)z = k_3(s_{\max} - x_P) \tag{22}$$

The flow equation of pump can be described as

$$Q_P = \frac{D_P}{2\pi}\omega_m = \frac{d^2}{8}z(s_{\max} - x_P)\omega_m = k_4(s_{\max} - x_P)\omega_m \tag{23}$$

The pump flow of inlet and outlet are

$$\begin{cases} Q_1 = Q_p - C_{ip}(P_1 - P_2) - C_{ep}(P_1 - P_{case}) \\ Q_2 = Q_p - C_{ip}(P_1 - P_2) - C_{ep}(P_2 - P_{case}) \end{cases} \tag{24}$$

where C_{ip}, C_{ep} are the internal and external leakage coefficient, P_{case} is the case pressure.

- **Cylinder model**

According to Newton's second law, the dynamic equation of the cylinder is

$$m_e\frac{d^2x_e}{dt^2} = A_e(P_1 - P_2) - B_e\frac{dx_e}{dt} - F_e \tag{25}$$

where A_e is the piston area, x_e is the displacement of EHA, A_e is area of EHA cylinder, m_e is the piston mass, B_e is the viscous damping coefficient.

Ignoring the leakage of hydraulic cylinder, the flow pressure equation between left and right chambers are

$$\begin{cases} \dot{P}_1 = \frac{\beta}{V_0 + A_e x_e}[Q_1 - A_e\dot{x}_e - C_{el}(P_1 - P_2)] \\ \dot{P}_2 = \frac{\beta}{V_0 - A_e x_e}[A_e\dot{x}_e + C_{el}(P_1 - P_2) - Q_2] \end{cases} \tag{26}$$

where C_{el} is the total leakage coefficient.

The block transfer function diagram of EHA is shown in Fig. 13 [22].

(3) **Mathematical model of HA/EHA**

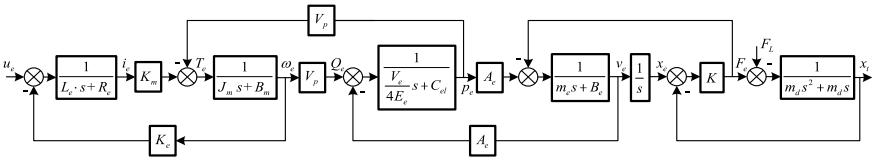


Fig. 13 The block transfer function diagram of EHA

HA adopts the proportional control, and EHA connects the two chambers to follow the HA as a damping force f_e. Considering that the equivalent mass and connection stiffness of a surface are m_d and K_t, respectively, the motion equation of the control surface can be written as

$$\begin{aligned} m_d \frac{d^2 x_t}{dt^2} &= F_h - F_e - F_L \\ F_h &= K_t(x_h - x_t) \\ F_e &= K_t(x_t - x_e) \\ F_e &= m_{pe} \frac{d^2 x_e}{dt^2} + B_{pe} \frac{dx_e}{dt} \end{aligned} \tag{27}$$

where m_{pe} is the mass of EHA and B_{pe} is the viscous damping coefficient of EHA.

Taking the Laplace transform of the Eq. (27), the control block diagram of the HA/EHA actuator is represented in Fig. 14.

(4) **Performance degradation of HA/EHA**

Here, taking the leakage of cylinder as example, the leakage due to wear and tear will increase when the cylinder operates shown in Fig. 15 [23].

The leakage flow can be expressed as [24]

$$Q = \frac{\pi d \delta^3 \Delta P}{12 \mu L C} \tag{28}$$

The leakage coefficient can be calculated as

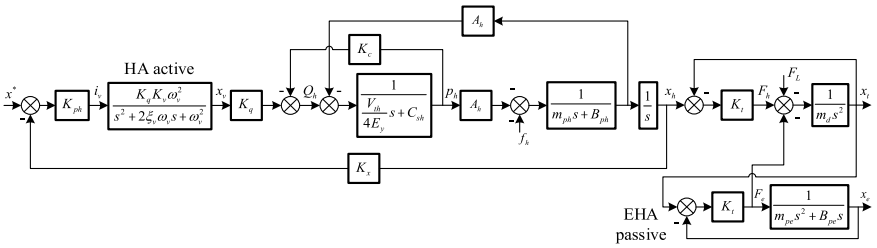


Fig. 14 System block diagram under HA active/EHA follower

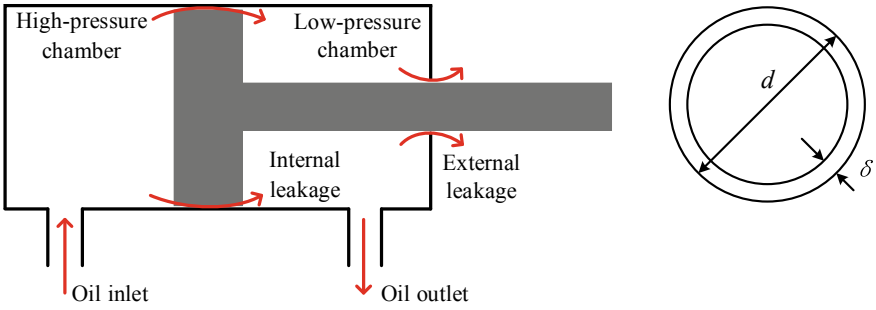


Fig. 15 The leakage diagram of a cylinder

$$C_{sh} = \frac{Q}{\Delta p} = \frac{\pi d \delta^3}{12 \mu L C} \tag{29}$$

where d is the piston diameter, δ is the single slit height, Δp is the pressure difference between two chambers, μ is the fluid dynamic viscosity, L is the travel distance of the piston, and C is the initial laminar correction coefficient.

With the operation of the cylinder, the wear volume between the cylinder and the piston can be described with the abrasive wear formula

$$\Delta V = K_s \frac{W L'}{H} \tag{30}$$

where ΔV is the wear volume, W is the normal load, H is the material hardness, K_s is the abrasive wear coefficient, and L' is the sliding distance.

Considering the number of actuation cycles n, the wear volume can be described as

$$\Delta V = K_s \frac{W n l_0 t}{H} \tag{31}$$

After replacing the wear volume into the leakage coefficient relationship, we obtain

$$C_{sh} = \frac{\pi d (\delta_0 + \Delta\delta)^3}{12 \mu L C} = \frac{\pi d \left(\delta_0 + \frac{K_s W n l_0}{H \pi d L} t\right)^3}{12 \mu L C} \tag{32}$$

where δ_0 is the initial slit height.

Considering the external load disturbance and wear unevenness, the leakage coefficient is subject to a normal distribution as follows

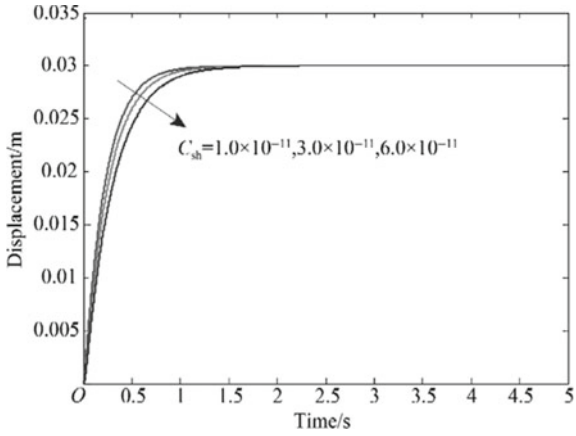


Fig. 16 The dynamic performance response under different leakage coefficients

$$f_{C_{sh}}(C_{sh},t)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\left[C_{sh}-\frac{\pi d\left(\delta_0+\frac{K_s W n l_0}{H\pi dL}t\right)^3}{12\mu LC}\right]^2}{2\sigma^2}} \tag{33}$$

The performance degradation curve can be obtained for different leakage coefficient shown in Fig. 16.

According to the mathematical model in Fig. 12, the system transfer function is

$$G(s)=\frac{A_h K_Q(s)K_t K_{ph}}{H(s)+A_h K_Q(s)K_{ph}G_{xF}(s)+G_{a4}(s)G_{xF}(s)} \tag{34}$$

Herein,

$$K_Q(s)=K_q\frac{K_v w_v^2}{s^2+2x_v w_v s+w_v^2} \tag{35}$$

$$\begin{aligned}H(s)=&\frac{V_{th}m_{ph}m_d}{4E_y}s^5+\left(\frac{V_{th}m_d B_{ph}}{4E_y}+K_{tm}m_{ph}m_d\right)s^4\\&+\left(K_{tm}B_{ph}m_d+A_h^2 m_d+\frac{V_{th}K_t m_d}{4E_y}+\frac{V_{th}m_{ph}K_t}{4E_y}\right)s^3\\&+\left(\frac{V_{th}K_t B_{ph}}{4E_y}+K_{tm}K_t m_{ph}+K_{tm}K_t m_d+A_h K_Q(s)K_{ph}m_d\right)s^2\\&+\left(K_{tm}B_{ph}K_t+K_t A_h^2\right)s+A_h K_Q(s)K_{ph}K_t\end{aligned} \tag{36}$$

$$G_{a4}(s)=\frac{V_{th}}{4E_y}m_{ph}s^3+\left(\frac{V_{th}}{4E_y}B_{ph}+K_{tm}m_{ph}\right)s^2$$

$$+\left(K_{tm}B_{ph}+A_h^2+\frac{V_{th}}{4E_y}K_t\right)s+K_{tm}K_t \tag{37}$$

$$G_{xF}(s)=\frac{K_t\left(m_{pe}s^2+B_{pe}s\right)}{\left(m_{pe}s^2+B_{pe}s+K_t\right)} \tag{38}$$

$$K_{tm}=K_c+C_{sh} \tag{39}$$

With the aforementioned transfer function, the system performance can be described as

$$T_r=G_{T_r}\left(K_j\right) \tag{40}$$

The parameter distribution is

$$\mathrm{K}_j\sim f_{K_j}\left(k_j,t\right) \tag{41}$$

The performance distribution can then be described as

$$T_r\sim g_{T_r}(t_r,t) \tag{42}$$

The performance reliability of actuation system can then be expressed as

$$P=\int_t^{\infty}\int_0^{v_{T_r}} g_{T_r}(t_r,t)dt_r dt \tag{43}$$

Because the mathematical model is very complicated, it is difficult to determine analytical solution to the performance reliability; therefore, one needs to resort to a numerical solution. The simulation flow chart shown in Fig. 17 can result in a solution to the performance reliability of the actuation system [24].

The reliability of the HA/EHA system depends not only on the performance degradation, but it can also be related to the critical component reliability. Assuming that the performance reliability is independent of the component reliability, we can define the integrated reliability of the actuation system as follows:

$$\begin{aligned}R(t,y)&=P\{\mathrm{Y}\in\Omega,T>t\}=P\{\mathrm{Y}\in\Omega|T>t\}\cdot P\{T>t\}\\R(t,y)&=R_T(t)\cdot R_Y(t)\end{aligned} \tag{44}$$

where the $R_T(t)$ is the component reliability. Because the main components of the actuation system consist of an amplifier, SV, cylinder, LVDT, and bypass valve, the component reliability is as follows:

$$R_T(t)=\prod_{i=1}^{5}R_T^i(t)=\prod_{i=1}^{5}e^{-\lambda_i t} \tag{45}$$

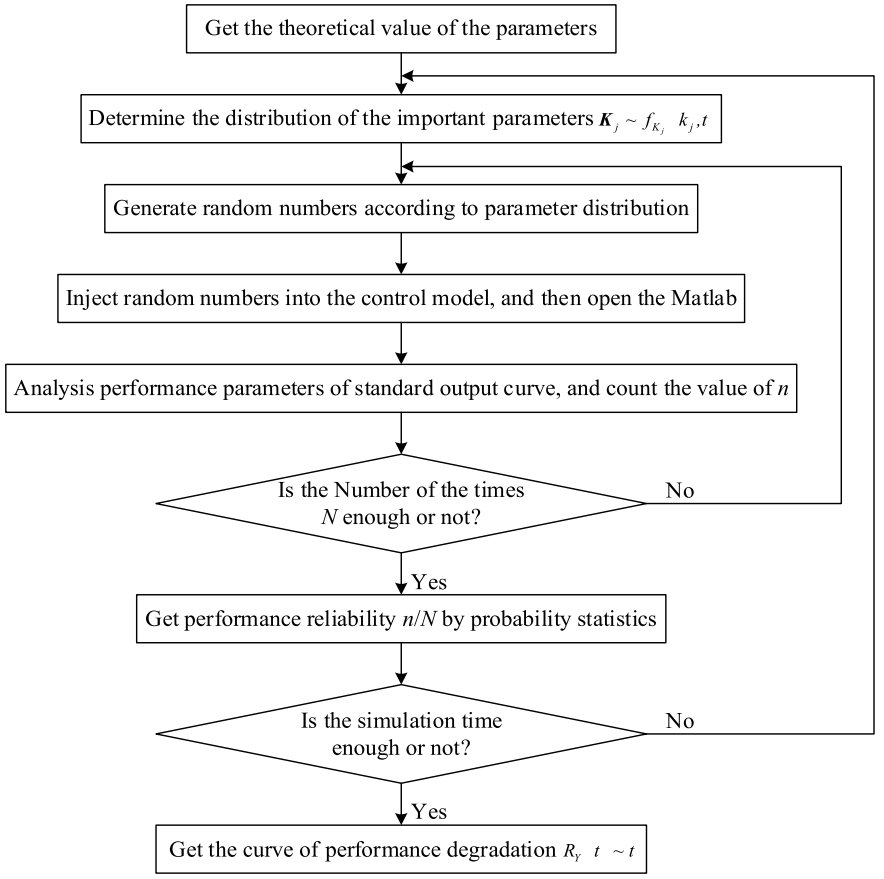


Fig. 17 Reliability simulation flow chart

where $R_T^i(t)$ is the reliability of components and λ_i is the component failure rate. The failure rate of individual components is listed in Table 4.

The reliability of the component is then $R_T(t) = e^{-0.00128t}$.

The required parameters to calculate the performance reliability are listed in Table 5.

Table 4 The component failure rate

Number	Component	Failure rate/per hour
1	Amplifier of servo valve	300×10^{-6}
2	Servo valve	360×10^{-6}
3	Cylinder	270×10^{-6}
4	LVDT	150×10^{-6}
5	Bypass valve	200×10^{-6}

Table 5 System Parameter Value

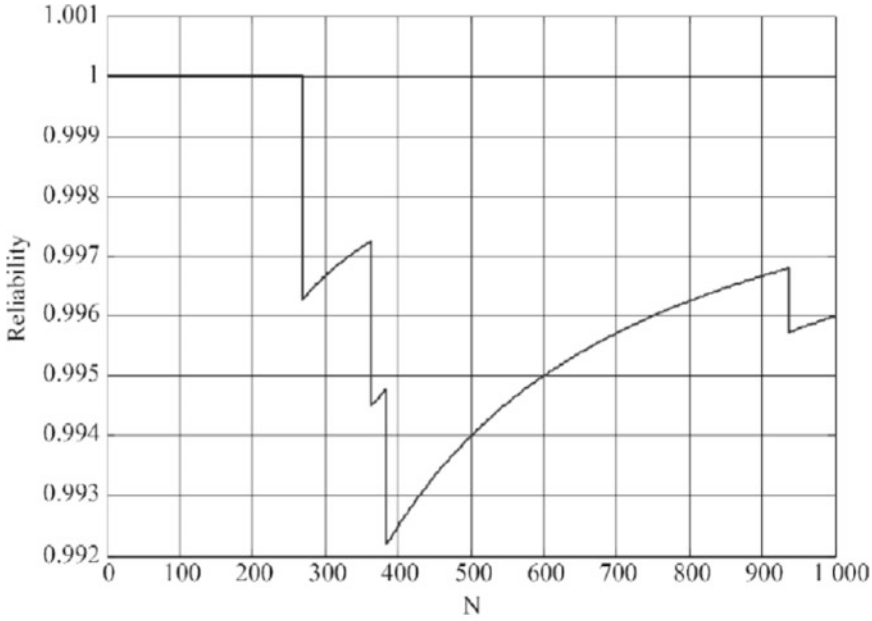
Parameter	Value	Unit	Parameter	Value	Unit
ξ_V	0.7	-	V_{th}	1.47×10^{-4}	m^3
ω_V	600	rad/s	E_y	8.0×10^8	Pa
K_V	1.52×10^{-4}	m/A	C_{sh}	1.0×10^{-11}	$(m^3/s)/Pa$
K_q	2.7	m^2/s	K_t	1×10^8	N/m
K_c	1.75×10^{-11}	$(m^3/s)/Pa$	m_d	600	kg
$m_{ph}(m_{pe})$	55	kg	$B_{ph}(B_{pe})$	10,000	Ns/m

Assuming the simulation number $N = 2000$, the time interval as $\Delta t = 0.5$h, the response time $T_r < 0.6$s, and the step input $x^* = 0.03$m, and substituting the parameters from Table 5 into the performance reliability, we obtain the performance reliability curve shown in Fig. 18.

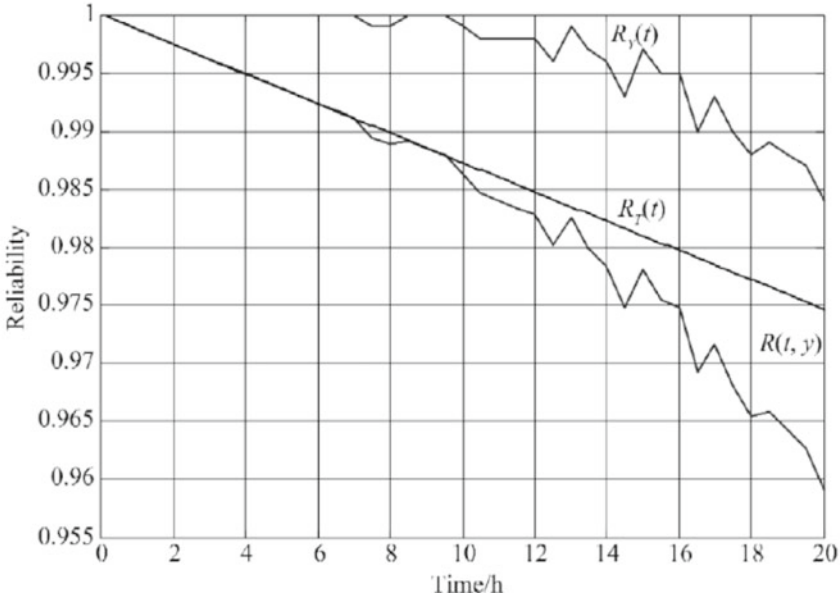
It is apparent that the integrated reliability takes into consideration the performance degradation with the system operation in addition to component reliability. Furthermore, increasing the simulation number improves the accuracy of the result and makes it closer to the real application.

4 Conclusions

This chapter summarizes the aircraft actuation system, explains the interface between the FCS and the actuation system. After introducing some typical actuation systems, the chapter provides some typical aircraft hydraulic actuation system constructions in current commercial aircraft. Furthermore, this chapter provides the comprehensive reliability definition and the reliability calculation methods. Through the redundancy and monitoring design, the system reliability can be improved. Afterward, the integrated reliability evaluation is provided for example of the actuation system. The results indicate that the system design meets the desired system reliability requirements.



(a) Performance reliability curve at $t = 14h$



(b) Integrated reliability curve under function and performance reliability

Fig. 18 Integrated reliability of an actuation system

References

1. Moir I, Seabridge A (2008) Aircraft systems: electrical and avionics subsystems integration. Wiley
2. McLean D (2000) Flight control systems: practical issues in design and implementation. In: Pratt RW (ed) The institution of electrical engineers, Michael Faraday House, Six Hills Way, Stevenage, Herts, SG1 2AY. 2000. 382 pp. Illustrated. £59.00. ISBN 0-85296-766-7; Aeronaut J 104(1037):334–335 (1968)
3. Crane D (2020) Dictionary of aeronautical terms
4. SAE International (2007) SAE AIR 5273: 2007 actuation system failure detection methods
5. Collinson RPG (2011) Aerodynamics and aircraft control. In: Introduction to avionics systems. Springer, Dordrecht
6. Bossche DV, Airbus (2006) The a 380 flight control electrohydrostatic actuators, achievements and lessons learnt
7. Wang Y (2014) Aircraft flight control hydraulic servo actuator. Aeronautic Industry Press, Beijing
8. Wang S, Tomovic M, Liu H (2015) Commercial aircraft hydraulic system. Shanghai Jiao Tong University Press
9. Frischemeier S (1997) Electrohydrostatic actuators for aircraft primary flight control—types, modelling and evaluation
10. Shi C, Wang S, Wang X, Wang J, Tomovic MM (2017) Active fault-tolerant control of dissimilar redundant actuation system based on performance degradation reference models. J Franklin Inst 354(2):1087–1108
11. Botten SL, Whitley CR, King AD (2000) Flight control actuation technology for next generation all-electric aircraft. Technol Rev J 8(2):55–68
12. Mei WH, Guo Y, Mang HY (2003) Practical counterexamples to AMSAA reliability growth model for multiple systems development in MIL-HDBK-189 and MIL-HDBK-781. In: Paper presented at the 8th ISSAT international conference on reliability and quality in design, proceedings
13. Cui X (2019) Energy based reliability modeling method and its application in aircraft flight control system, Beihang University
14. Li T, Wang S, Shi J, Zio E, Cui X (2020) An energy-based coupling degradation propagation model and its application to aviation actuation system. Chin J Aeronaut 33(4):1288–1298
15. Kołowrocki K, Soszyńska-Budny J (2011) Modeling reliability and safety of multistate systems with ageing components. In: Reliability and safety of complex technical systems and processes. Springer Series in Reliability Engineering. Springer, London
16. Wang S (2000) Reliability engineering. Beijing University of Aeronautics and Astronautics Press
17. MIL-F—9490D (USAF) Military specification, flight control systems—design, installation and test of piloted aircraft, general specification for, June 6, 1975
18. Wang S, Cui X, Shi J, Tomovic MM, Jiao Z (2016) Modeling of reliability and performance assessment of a dissimilar redundancy actuation system with failure monitoring. Chin J Aeronaut 29(3):799–813
19. Cui X, Wang S, Li T, Shi J (2019) System reliability assessment based on energy dissipation: modeling and application in electro-hydrostatic actuation system. Energies 12(18):3572
20. Shi C (2018) Research on variable load failure mechanism and life prediction of load sensing electro-hydrostatic actuator, Beihang University
21. Shi C, Wang S, Wang X, Zhang Y (2018) Variable load failure mechanism for high-speed load sensing electro-hydrostatic actuator pump of aircraft. Chin J Aeronaut 31(5):949–964
22. Wang S, Tomovic MM, Liu H (2016) Chapter 4—new technology of aircraft hydraulic system

23. Yin Y (2022) Vector engine control system fault coupling mechanism and information fusion fault diagnosis method, Beihang University
24. Li Q, Wang S, Shi J, Wang X, Jiao Z (2014) Reliability modeling analysis for hydraulic/electro-hydrostatic dual redundant actuation system. In: Paper presented at the proceedings of 2014 IEEE Chinese guidance, navigation and control conference

Integration of Reliability Design, Installed Base, and After-Sales Services for System Availability

Tongdan Jin, Wenjin Zhu, and Ziqi Jiang

Abstract System availability is a fundamental measure to evaluate the reliability performance of capital goods. Traditional approaches to availability management, such as reliability-redundancy allocation, preventive maintenance, and spare parts logistics, usually focus on a particular phase of system life. This paper discusses a holistic lifetime approach to sustaining system availability in an integrated product-service framework. Our approach seamlessly incorporates reliability, redundancy, maintenance, repairable inventory, and installed base information into a unified availability measure. A superimposed renewal process is adopted to characterize spare part demands considering the effect of installed base and proactive replacements. Extensive simulations are conducted to analyze the spares demand profile in terms of maintenance time, lifetime distribution, inventory lead time, and repair and renewing capacity. The study reveals that: (1) system availability is jointly determined by ten performance drivers across the product design, manufacturing, and after-sales market; (2) Poisson spare parts demand assumption is valid provided the item lifetime is much longer than the inventory replenishment time; and (3) installed base information provides a causal approach for spares demand forecasting during the new product introduction phase.

Keywords Superimposed renewal process · Installed base · Erlang queue · Non-homogeneous Poisson process · Reliability-maintenance-inventory integration

T. Jin
Ingram School of Engineering, Texas State University, San Marcos, TX 78666, USA

W. Zhu (✉)
School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China
e-mail: wenjin.zhu@nwpu.edu.cn

Z. Jiang
Department of Systems Engineering, Cornell University, Ithaca, NY 14853, USA

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_10

1 Introduction

High system availability is essential to the daily operations in private and public sectors, such as manufacturing, transportation, telecommunication, energy generation, and homeland security, among others. Downtime of capital goods often results in staggering amounts of production loss or even risk to human life. For instance, when an A330-200 plane is grounded for a routine maintenance, it usually takes 3 days and consumes 150 man-hours. The total fixed cost of maintenance can be as high as $67,000, including aircraft leasing costs, labor hours, and spare parts. The opportunity cost because of 3-day ground may reach $700,000 assuming the plane's revenue is $50 per passenger-hour. To ensure system availability, the original equipment manufacturers (OEM) strive to design reliable products as well as to provide responsive maintenance, repair and overhaul (MRO) services. If the system is down, MRO is expected to be carried out promptly to reduce the downtime. We focus on the integrated OEM who designs and produces capital equipment and also provide repair and spare parts in the after-sales market. In literature, the strategy unifying product design, manufacturing, and sustainment is also referred to as an integrated product-service system (PSS) or integrated product-service offering (PSO) [1–3]. Throughout the paper, equipment, product and machine are used interchangeably, representing a repairable system comprised of multiple items or parts.

Three approaches have been frequently discussed in literature with the goal of achieving high system reliability and availability performance. These are reliability-redundancy allocation (RRA), preventative maintenance (PM), and spare parts logistics (SPL). Albeit the effectiveness, existing methods usually focus on certain period of the product lifetime. For instance, RRA usually concentrates on product design and manufacturing phase, while PM and SPL are confined to the post installation. In fact, reliability and availability performance is dependent upon design method, production technology, maintenance policy, repair capacity, and spares inventory. Hence it is imperative to seek a holistic management solution throughout the product's lifetime. To that end, this paper presents an integrated reliability-maintenance-inventory (RMI) approach to manage the availability of repairable systems through the synthesis of RRA, PM, SPL and installed base information. Particularly, it aims to guide the OEM to allocate reliability, redundancy, maintenance, spare parts, and repair resources to achieve high system availability goal at a low cost.

Our study contributes to reliability modeling and service logistics management in three aspects. First, the RMI approach represents a first-of-its-kind by integrating reliability, redundancy, maintenance, spares inventory, and repair capacity in a unified framework. This type of product-service integration model has rarely been discussed in reliability engineering and operations management literature. Second, the model can handle both homogeneous and non-homogenous Poisson spares demands. For Poisson demand, two parallel Erlang-C queues are adopted to cope with failure and planned replacements, respectively. The queues can differentiate the levels of repair skillset, cost, and hands-on time of two types of replacement demands. For non-homogenous Poisson demand, repair and renewing turn-around time is approximated

based on the coefficient of variation of parts arrival time and service time. Third, over 500,000 simulations covering a large scope of parameter values are conducted to demonstrate the applicability of the proposed RMI approach.

The remainder of the paper is organized as follows. Section 2 reviews existing RRA, PM and SPL models. Section 3 discusses the needs of implementing an integrated reliability-maintenance-inventory solution for achieving high system availability. Section 4 estimates the failure and the planned replacements for a stationary system fleet based on superimposed renewal theory. Section 5 uses the Boeing 787 fleet to characterize the lead time spares demand under a growing installed base. In Sect. 6 simulation programs are developed to validate the lead time Poisson spares demand. Section 7 presents the system availability model with the integration of RRA, PM and SPL. Section 8 concludes the work. Table 1 lists the parameters of the integrated system availability model.

2 The State-of-the-Art

A large body of literature is available pertaining to the optimization of RRA, PM and SPL at the component and system level. In this study, component, part, item, and modules are used interchangeably, representing a line replaceable unit (LRU) of a system. In RRA, the system's mean-time-between-failures (MTBF) can be improved through the adoption of advanced design, durable materials, or redundant components subject to resources or cost constraints [4]. An early review of RRA models is available in the work of [5], and two recent ones are referred to [6] and [7]. Since the RRA problem turns out to be an NP-hard issue, various solution techniques have been proposed, including branch-and-bound [8], genetic algorithm [9], Tabu search [10], artificial bee colony [11], multi-objective [12], Pareto optimality [13], importance measure [14], among others. The majority of RRA models concentrate on the product design and manufacturing phase, while maintenance and repair in the after-sales market are not explicitly considered.

The purpose of PM is to periodically inspect the system and proactively replace the parts prior to the incipient failure. For instance, in age- or usage-based maintenance, an LRU is replaced either it reaches a predefined threshold or fails randomly [15]. Flexible age replacement policy has been investigated as well, and the pros and cons are compared with fixed age replacement policy [16–18]. With the wide use of sensor technology, predictive maintenance or condition-based maintenance (CBM) has been gaining growing attention. In CBM, the health condition or degradation precursor (e.g., electrical, thermal and mechanical signals) is monitored via in-situ sensors. The remaining useful life (RUL) is predicted through a diagnostic and prognostic health management (PHM) program. The replacement action is taken when the RUL approaches, but not exceeds, a pre-defined threshold. PHM program is typically established upon statistical model, Bayesian inference, machine learning algorithms, or other analytics methods [19, 20]. A dynamic maintenance policy integrates periodic preventive maintenance, reactive maintenance and opportunity maintenance is

Table 1 Notation of model parameters

Notation	Comment
A	Availability of single-item system
R(t)	Reliability function of line replaceable unit (LRU)
F(t)	Cumulative distribution function
M(t)	Number of renewals or replacements during [0, t]
N(t)	Number of systems installed during [0, t]
S(t)	Cumulative spares demand of a fleet in [0, t]
D(t)	Inventory lead time demand during [t, t + l]
T_p	Mean downtime in a planned replacement
T_q	Mean downtime in a failure replacement
T_{MTBR}	Mean time between replacements of LRU
T_{MTBF}	Mean time between failures of LRU
B_q	Probability for a part waiting in a repair queue
B_p	Probability for a part waiting in a renewing queue
W_i	Installation time of system i for i = 1, 2, …, N(t)
f(t)	Probability density function of LRU lifetime
α, β	Weibull scale and shape parameters, respectively
λ_p	Aggregate fleet part demand rate of planned replacement
λ_q	Aggregate fleet part demand rate of failure replacement
ρ_p, ρ_q	Renewing and repair traffic intensity rate, respectively
μ_p, μ_q	Part renewing rate and repair rate, respectively
k	Minimum required working item in a system
l	Lead time for replenishing the spares inventory
m	System fleet size or installed base
n	Total number of components in a system
t_p	Part renewing turn-around time
t_q	Part repair turn-around time
t_s	Hands-on time for replacing a part
s	Base-stock level of spare parts for LRU
τ	Maintenance interval for LRU
p	Number of renewing channels or servers for LRU
q	Number of repair channels of servers for LRU

proposed by [21]. Reviews on the state-of-the-art of PM and CBM can be found in the works by [22–24].

SPL aims to reduce the equipment downtime by promptly supplying a spare item to replace the failed unit through a service supply chain. The goal of the SPL model is to locate the stockrooms and allocate the spares inventory to meet the intermittent demand generated by geographically dispersed customers [25, 26]. In the SPL model, trade-offs must be made between the cost and the supply chain variables, such as the inventory level, the stock location, and transportation time. Typical performance measures are fill rate, backorders, parts availability, inventory capital, and holding cost [27–29]. Recently, much attention is paid to jointly allocating spares inventory, maintenance policy, and repair capacity because many RLU can be renewed, refurbished, and reused. Along this line, De Smidt-Destombes et al. [30] jointly optimize the maintenance initiation, the spare parts, and the repair capacity to minimize the ownership cost of a single k-out-of-n system. Jin et al. [31] propose a principal-agent model that jointly allocates the maintenance time, spares inventory, and parts repair and renewing capacity for minimizing the annualized fleet cost of k-out-of-n systems. Basten and Ryan [32] investigate the impact of delay in performing planned replacement on the optimal spares inventory policy. Zhu et al. [33] consider the impact of logistic delay when optimizing the maintenance policy for a wind turbine. An early survey on SPL models was made by Basten and van Houtum [34], and a more recent one is available in the work of [35].

Although various models pertaining to RRA, PM and SPL have been developed to achieve the system reliability and availability goal, the majority are focused on a particular phase of the product lifetime. There is a lack of lifetime availability approach in which reliability, redundancy, maintenance, spares, repair, and installed base are jointly considered. This study aims to incorporate RRA, PM, SPL, and fleet size into a unified system availability framework for both stationary and non-stationary installed bases.

3 Reliability-Maintenance-Inventory Integration

3.1 Best Practice of Product-Service Offering

To understand the motivation of reliability-maintenance-inventory integration, we present an industry case to show how the product and the service are seamlessly integrated to attain the system availability goal of field installation. Automated test equipment (ATE) is a capital-intensive machine widely used for wafer probing and device testing in semiconductor manufacturing plants. Users of ATE include Intel, IBM, Texas Instruments, Samsung, Philips, TSMC, among others. The cost of an ATE machine varies between $1 million to $3 million depending on the equipment configuration and performance. Modularity design is adopted to facilitate the system maintenance and repair tasks. Due to the complexity of technology, the OEM is

responsible for the ATE design, manufacturing, maintenance, repair, and spare parts supply. Upon failure, the bad module is swapped with a spare item so that the machine can be restored to the production promptly. The bad module is returned to the repair center for root-cause analysis. After fixing, it is returned to the spares inventory for future use.

Figure 1 shows how the OEM leverages global resources to design, manufacture and sustain the ATE fleet in various customer sites. The product-service supply chain spans across Asian Pacific, North America, South America, Europe, and Middle East regions. For instance, the ATE design is done in Boston, MA and San Jose, CA, USA. The software and application programs are outsourced to the subcontractors in India. The printed circuit boards (i.e., the LRU) are made in Charlotte, NC, USA and then shipped to the factory in Shanghai, China where the ATE system is assembled. The assembled machine is then shipped to different customers in the world.

To ensure system availability, spare parts inventory is co-located at the customer sites. As shown in Fig. 1, spares stockrooms are established in Europe, East Asia, and North America where large amounts of systems are installed. To provide 24/7 repair services, two repair crews are established in the Philippines and Costa Rica, respectively. Both countries are chosen because of low labor cost and less stringent labor laws. All repaired or renewed parts are shipped to the central warehouse in Memphis, TN, from where spare parts are further distributed to regional stockrooms.

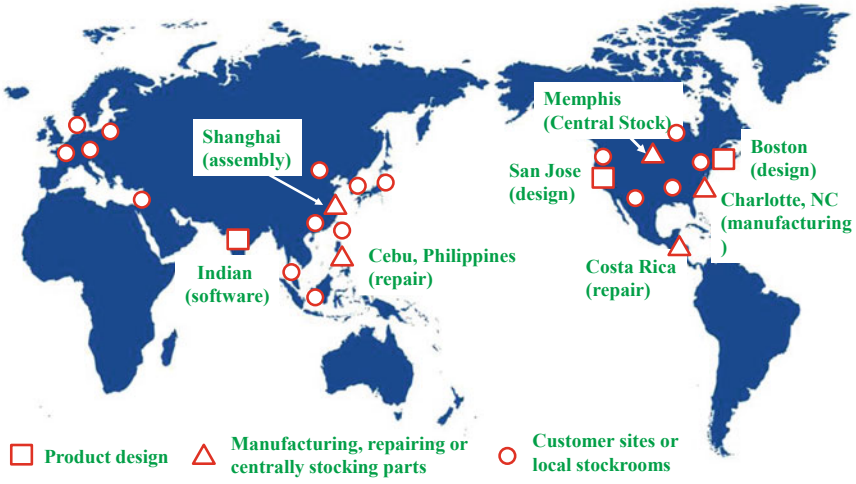


Fig. 1 The integrated product-service supply chain of ATE industry [36]

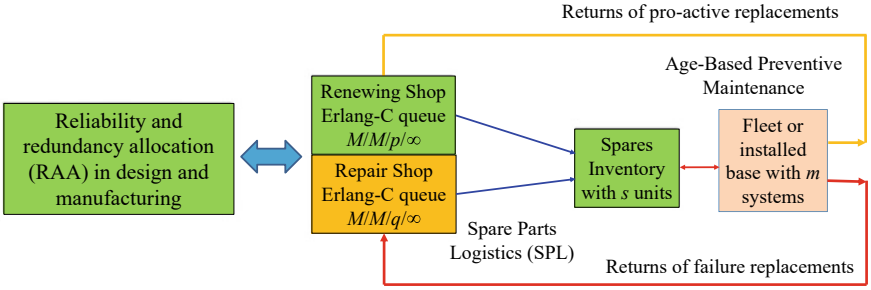


Fig. 2 Single echelon supply chain for system availability support

3.2 Modeling Reliability-Maintenance-Inventory Integration

Without loss of generality, Fig. 2 shows a single-echelon supply chain that offers integrated product-service capability. In this setting, a field fleet comprised of m systems is sustained under age-based maintenance. Critical items within the system are inspected at a pre-defined time interval τ. A spares inventory is located in proximity to the system fleet to facilitate corrective and pro-active replacements.

If the item survives and reaches τ, it is pro-actively replaced with a good part supplied by the local stockroom. The replaced item, though functional, is sent back to the repair shop for renewal. If the item fails prior to τ, a corrective replacement is performed, and the failed unit is sent back to the repair shop for troubleshooting.

Queueing model has long been used to characterize the performance of repairable inventory systems in literature [27, 37]. Poisson demand is usually assumed in the repairable inventory literature because of mathematical tractability and technical applicability. Figure 3 graphically shows the working principle of the part repair and renewal processes. Since less recourse and time are consumed in renewing a degraded part than repairing a failed unit, two separate queues are adopted to model the renewing and the repair jobs. Particularly, we use $M/M/q/\infty$ to characterize the repair process, and $M/M/p/\infty$ queue to characterize the renewing process. Note that q and p are the number of repair and renewing servers, respectively.

3.3 Repair and Renewing Queues Under Poisson Demand

The $M/M/q/\infty$ queue is also known as the Erlang-C model. It is appropriate to characterize repairable inventory system as it accommodates a waiting line when all repair servers are busy. For an Erlang-C repair queue, the probability that an incoming part needs to wait is given as follows [38]:

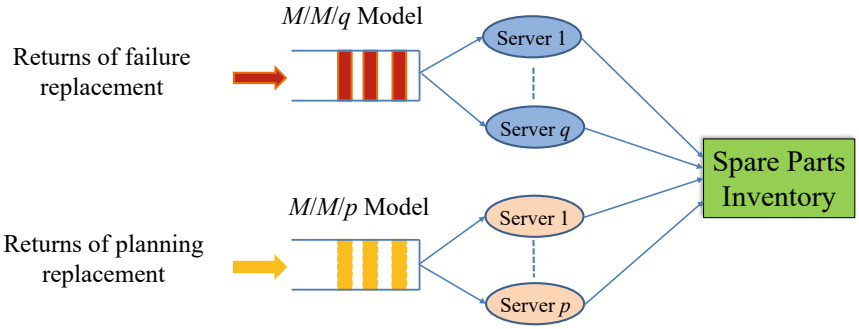


Fig. 3 Working principle of repair and renewing of field returns

$$B_q = \frac{(q\rho_q)^q}{q!(1-\rho_q)}\left[\sum_{j=0}^{q-1}\frac{(q\rho_q)^j}{j!} + \frac{(q\rho_q)^q}{q!(1-\rho_q)}\right]^{-1} \tag{1}$$

where λ_q is the part arrival rate, and μ_q is the part repair rate per server. Note $\rho_q = \lambda_q/(q\mu_q)$ is called the repair traffic intensity rate. The repair shop queue is stable if and only if $\rho_q < 1$.

Figure 4 depicts the relation between q and B_q at different ratios of λ_q and μ_q. The lines from the left to the right, represent $\lambda_q/\mu_q = 0.5$, 0.9, 1.5, 2, and 3, respectively. It shows the waiting probability quickly decreases as the repair servers increase. For instance, given $\lambda_q/\mu_q = 0.5$, the waiting probability is 0.5 for $q = 1$, and it drops to 0.1 for $q = 2$. Similarly, given $\lambda_q/\mu_q = 0.9$, the waiting probability is 0.9 for $q = 1$, and it drops to 0.28 for $q = 2$. If $\lambda_q/\mu_q \geq 1$, at least two servers are required to maintain the queue stability.

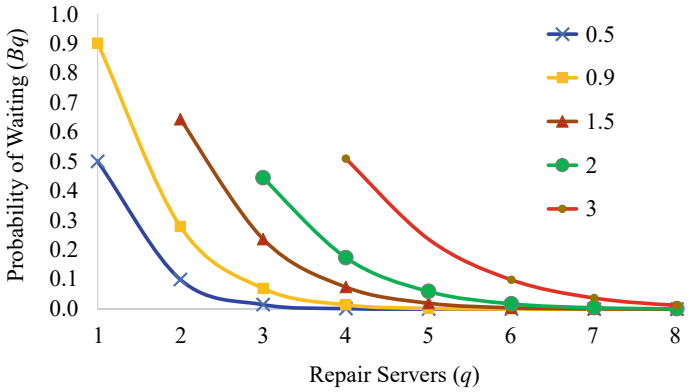


Fig. 4 Waiting probability versus the number of the repair servers

When the transportation time between the repair shop and the spares inventory is small compared to the repair time, the repair turn-around time (TAT), denoted as t_q, is the sum of the part waiting time and the actual hands-on repair time. That is

$$t_q = \frac{B_q}{q\mu_q - \lambda_q} + \frac{1}{\mu_q} \tag{2}$$

Similarly, we can analyze the part renewal process using $M/M/p/\infty$ queue where p is the number of renewal servers. The probability that an incoming part needs to wait prior to being renewed is given by

$$B_p = \frac{(p\rho_p)^p}{p!(1-\rho_p)}\left[\sum_{j=0}^{p-1}\frac{(p\rho_p)^j}{j!} + \frac{(p\rho_p)^p}{p!(1-\rho_p)}\right]^{-1} \tag{3}$$

where λ_p is the renewing part arrival rate, and μ_p is the renewing rate per server. Note that $\rho_p = \lambda_p/(p\mu_p)$ is called the renewing traffic intensity rate. The renewing queue is stable if and only if $\rho_p < 1$. The renewing turn-around time, denoted as t_p, is the sum of the part waiting time and the hands-on renewing time. That is

$$t_p = \frac{B_p}{p\mu_p - \lambda_p} + \frac{1}{\mu_p} \tag{4}$$

3.4 Repair and Renewing Queues for Non-Markovian Process

One of the key characteristics of homogenous Poisson process (HPP) is that the coefficient of variation (CV) is equal to one. CV is defined as the ratio of the standard deviation to the mean of the population. The higher the CV, the greater the dispersion of the demand. If CV < 1, Eq. (2) or (4) tends to overestimate the delays while the opposite is true if CV > 1. Since exact or closed-form solution for non-Markovian multi-server queues is still not available, several good approximations have been proposed [38, 39]. Let $t_{q,G}$ be the part repair TAT with general or non-Markovian arrival process, then we have

$$t_{q,G} = t_q\left(CV_{aq}^2 + CV_{sq}^2\right)/2 \tag{5}$$

where ${CV_{aq}}^2$ and ${CV_{sq}}^2$ are the square of the CV of the part inter-arrival time, and the part repair time, respectively. Note that t_q is the repair TAT given in Eq. (2). Similarly, let $t_{p,G}$ be the part renewing TAT in a non-Markovian queue. Then we have

$$t_{p,G} = t_p\left(CV_{ap}^2 + CV_{sp}^2\right)/2 \tag{6}$$

where $CV_{ap}{}^2$ and $CV_{sp}{}^2$ are the square of the CV of the part inter-arrival time and renewing time, respectively. Note that t_p is the TAT given in Eq. (4).

4 Spares Demand Under Preventive Maintenance

4.1 Aggregate Fleet Spare Parts Demand

The demand formed by the union of item replacements of a fleet is called the superimposed renewal process (SRP). Cox and Smith [40] have proved that, as the fleet size approaches infinity, the SRP becomes an HPP regardless of the lifetime distribution of the items. Wang [41] further shows that the occurrence times between two successive replacements can be approximated exponentially if: (1) the fleet size is at least ten, and (2) item failures are mutually independent. Both conditions are generally satisfied in real-world applications where a group of machines of the same type are used. For instance, The US Southwest Airlines has nearly 750 aircraft, and they all belong to the Boeing 737 family [42]. More discussions on SRP and its variants can be referred to [43].

Two spares demand streams are generated under PM. One is due to the failure replacement, and the other is due to the planned replacement. Figure 5 shows how the two demand streams are generated from m systems under age-based PM. Without loss of generality, each system contains one LRU only. If the system fails prior to τ, a corrective replacement is performed; otherwise, a pro-active replacement is made at τ. All systems independently generate corrective or pro-active replacements over time. The aggregate fleet failures are the superposition of the corrective replacements of m systems. Similarly, the aggregate fleet planned replacements are the superposition of the planned replacements of m systems.

4.2 Mean Time Between Consecutive Failures

It is worth mentioning that the HPP spares demand model in Sect. 4.1 is established upon corrective maintenance. New theoretical exploration is needed to show how consecutive failure replacements behave under age-based PM. Let Y be the time between two consecutive failure replacements. Then Y is a random variable and can be represented by

$$Y = N\tau + T \tag{7}$$

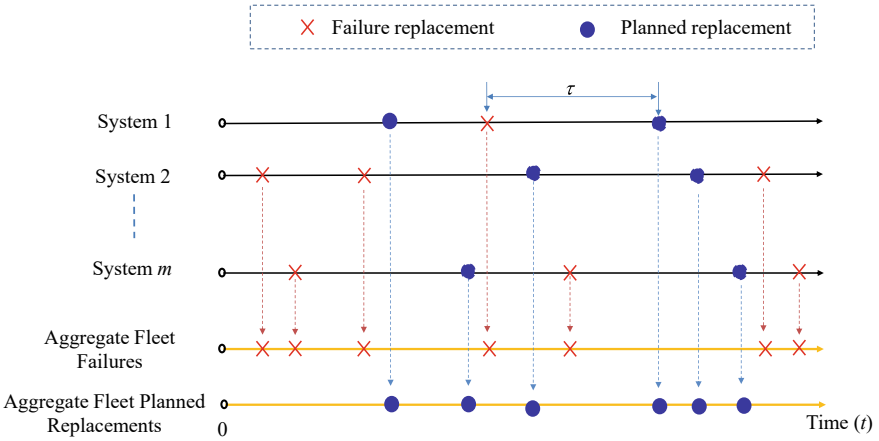


Fig. 5 Aggregate replacements for a system fleet under PM

where T is the time-to-failure after N planned replacements. The random number N has a geometric distribution and its expected value is given by

$$E[N] = \sum_{n=0}^{\infty} n \Pr\{N = n\} = \frac{R(\tau)}{F(\tau)} \tag{8}$$

Now $E[Y]$ can be obtained as follows

$$E[Y] = E[N]\tau + E[T] = \frac{\tau R(\tau)}{F(\tau)} + \frac{1}{F(\tau)} \int_0^{\tau} (F(\tau) - F(t))dt = \frac{1}{F(\tau)} \int_0^{\tau} R(t)dt \tag{9}$$

In defense and military logistics literature, $E[Y]$ is also referred to as the mean time between unscheduled removals.

4.3 Mean Time Between Consecutive Planned Replacements

Under age-based PM, the time between two consecutive planned replacements is also random. Let Z be the time between two consecutive planned replacements. Then Z is a random variable that can be estimated as

$$Z = MT + \tau \tag{10}$$

where T is the time-to-failure of the system, and M is a random number representing the consecutive failure replacements prior to the last planned replacement. Note that

M follows the geometric distribution, and the expectation is obtained as follows

$$E[M] = \sum_{m=0}^{\infty} m \Pr\{M = m\} = \frac{F(\tau)}{R(\tau)} \tag{11}$$

Finally, the expectation of Z is obtained as follow

$$E[Z] = E[M]E[T] + \tau = \frac{1}{R(\tau)} \int_0^{\tau} R(t)dt \tag{12}$$

Equations Eq. (9) and (12) show that Y and Z depends on τ and $R(t)$ or $F(t)$. More detailed discussions can be referred to [31].

4.4 Demand Characterization Using Simulation

Besides the analytical approach, simulation and design of experiments are also the effective methods to investigate the spare parts demand behavior. This section presents the simulation result of spares demand under PM originally reported in [44]. Assuming Weibull lifetime distribution, the key parameters include Weibull scale and shape parameters α and β, fleet size m, maintenance interval τ, planning horizon H and system installation time W. Without loss of generality, we set $\alpha = 1$. The simulation is conducted on a fleet of single-item systems, but the result is applicable to muti-item systems. Table 2 summarizes the range of the parameters in the simulation.

A total of 14 cases are investigated. Cases 1–12 assume all systems are installed at $W = 0$. For Cases 13 and 14, the system installation time is uniformly distributed between $[0, 0.5T_{MTBF}]$ where T_{MTBF} is the system mean-time-between-failures. In a case such as $\beta = 1.5, m = 5, \tau = 0.5\,T_{MTBF}, H = 5T_{MTBF}$, and $W = 0$, the inter-arrival times of spare demands of the fleet are simulated and recorded for both failure and planned replacements, respectively. The inter-replacement times are further used for exponential fitting. Two observations are made by [44]. First, among all the cases, the aggregate fleet failures follow an HPP independent of W. Second, the aggregate fleet planned replacements tend to be HPP for $W = 0$. However, the planned replacements tend to be an NHPP if W is uniformly distributed. Thus, Eqs. (5) and (6) shall be used to estimate the repair and renewing TAT under NHPP, respectively.

Table 2 Key parameters in simulation under PM with $\alpha = 1$

Range	β	m	τ	H	W
Lower limit	1.5	5	$0.5T_{MTBF}$	$5T_{MTBF}$	0
Upper limit	5	20	$2T_{MTBF}$	$30T_{MTBF}$	$0.5T_{MTBF}$

5 Spare Parts Demand Under a Growing Installed Base

5.1 Relation Between Parts Demand and Installed Base

The demand analysis thus far is applicable to a system fleet with a stable size. This is typically the case when the market enters the mature phase. During the introduction of a new product, the fleet size or the installed base keeps growing due to the market expansion. This process may last for years or even several decades depending on the nature of the product. For instance, car manufacturers usually introduce a new model every 2–4 years, therefore, the fleet size of the new model keeps growing during the introduction period. For wind turbines and airplanes, the fleet size of a particular model will grow for 10–20 years. Figure 6 shows Boeing delivered the first three B787 planes in 2011, the installed base reached 1008 by 2022 [45]. The annual installation varies significantly between 2011 and 2022. The largest delivery occurred in 2019 with 158 new planes shipped. On average, each year 84 new planes are added to the existing fleet with a standard deviation of 60 planes.

As the fleet size expands over time, the spares inventory must operate dynamically and possess sufficient parts to handle the time-varying demand rate. The existing fleet continues to generate the demand for spare parts because of corrective and planned replacements. On the other hand, since new systems are being added, more spare parts are needed on top of the current installed base. Thus, the spares demands are generated by the superposition of the existing systems and the newly installed systems. In fact, the spares demand turns out to be a non-stationary process with increasing mean and variance under a growing installed base.

Figure 7 shows how the number of spare engines goes up per annual to meet the preventive maintenance as the Boeing 787 fleet size expanded between 2011 and 2022. Assuming an engine requires an overhaul service after accumulating

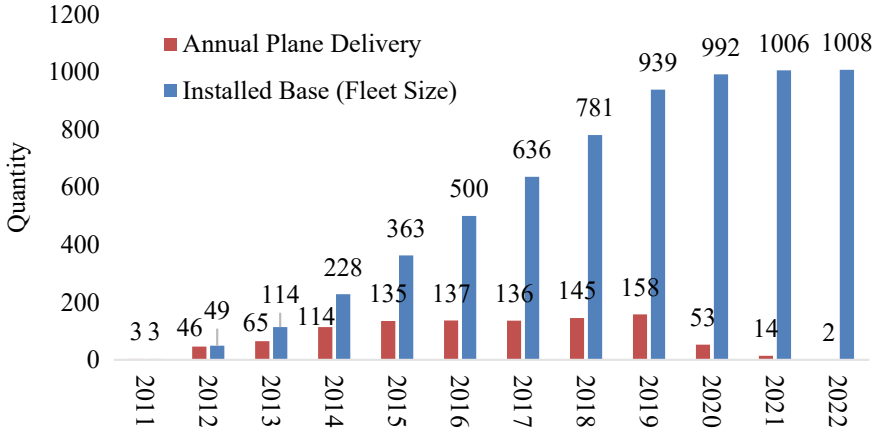


Fig. 6 Annual plane delivery and cumulative installed base of Boeing 787

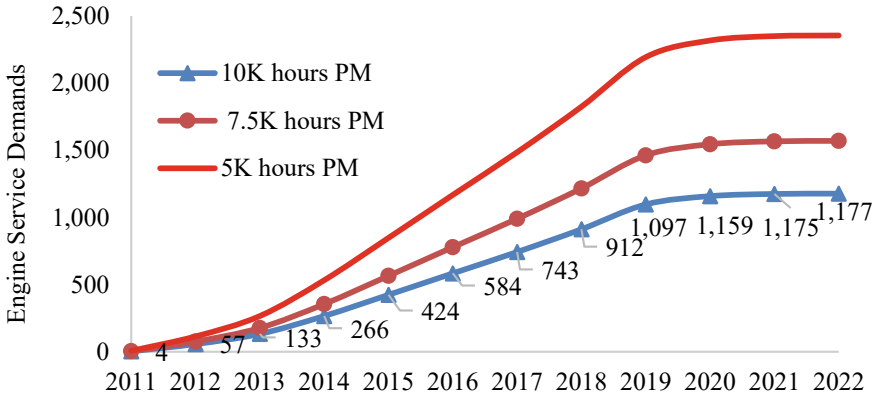


Fig. 7 Engine service demands under a growing 787 fleet

10,000 flight hours. In 2011, only 4 engines required the overhaul. This estimation is made assuming a plane is powered by two engines and operates 16 h per day. The engine demand increases with the installed base of the aircraft. By 2022, there are 1177 engines requiring overhaul services as the fleet size reaches 1008. Similar observations can be made when the engine PM interval becomes 5000 and 7500 h.

If an engine needs to be overhauled, it can be removed and replaced with a spare engine. This enables the plane to continue the service without an extended period of grounding for maintenance. Assume the engine's overhaul turn-around time is one week. When the fleet size is small, a small number of spare engines are needed to meet the maintenance demand. For instance, in 2013 the inventory needs to provide three spare engines under the 10,000-h PM policy. However, the spare engine stock level goes up with the growing installed base. In 2022, when the number of planes in service reaches 1008, the inventory must hold at least 23 spare engines so that the overhaul demands can be satisfied. This represents a huge amount of asset cost. The Rolls-Royce engine, one of two engines for the Boeing 787 jetliner, costs about $20 million at list price [46]. The cost of purchasing and stocking 23 spare engines will reach $460 million. If the engine's maintenance interval is reduced from 10,000 to 5000 h, the inventory level almost doubles and 45 spare units are needed to meet the overhaul demand as shown in Fig. 8. As a result, the total inventory asset cost increases to $900 million.

5.2 Cumulative Spare Parts Demand Model

Since spares demand is correlated with the fleet size, it is imperative to incorporate the installed base information into the demand forecasting. Below we present a spares demand forecasting model that synthesizes the renewal integral equation with the uncertain growth of the installed base. This model was originally presented in the

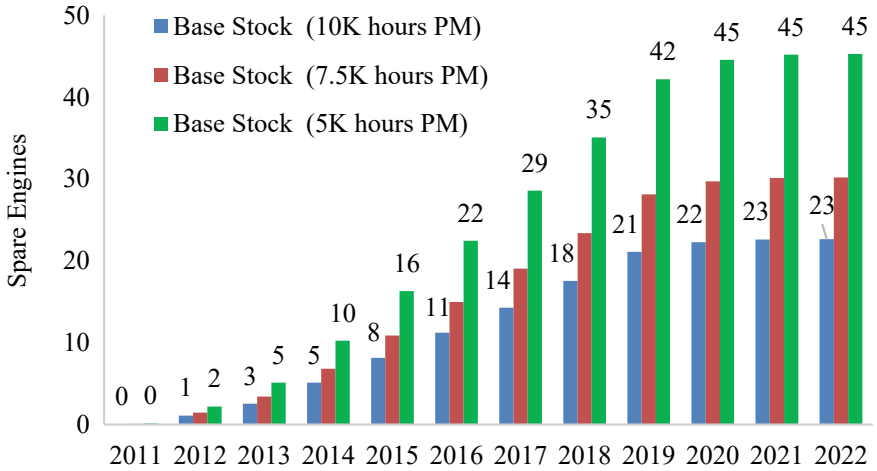


Fig. 8 Base stock level of spare engines under a growing aircraft fleet

work of [47]. Assume the first system is installed at time 0. There will be $N(t) + 1$ systems installed in $[0, t]$, where $N(t)$ is a random variable representing the systems installed in $(0, t]$. Under the corrective maintenance, the cumulative spare parts demand of the fleet in $[0, t]$, denoted as $S(t)$, can be estimated as

$$S(t) = M(t) + \sum_{i=1}^{N(t)} M(t - W_i), \quad \text{for } t \geq 0 \tag{13}$$

where

$$M(t) = F(t) + \int_0^t M(t - x) dF(x) \tag{14}$$

Here $M(t)$ is the number of replacements during $[0, t]$ for the system installed at time 0. Similarly, $M(t - W_i)$ is the number of replacements during $[W_i, t]$ for the ith system installed at $W_i > 0$. Note that W_i is a random variable because customer's decision on purchasing a new system is unknown. Equation (14) is called the renewal-integral equation, and $F(t)$ is the cumulative distribution function of the system lifetime. $M(t)$ is rather difficult to be solved for general lifetime distributions. An explicit solution for $M(t)$ is available only for certain distributions, such as exponential or Erlang lifetime.

5.3 Lead Time Spare Parts Demand Model

The spare parts demand during the inventory replenishment lead time is of particular interest to a maintenance planner. Lead time is the time that elapses from when the order is placed to when the part is received by the inventory. Let l be the replenishment lead time. Then the spares demand during $[t - l, t]$, denoted as $D(t)$, is obtained as follows.

$$D(t) = S(t) - S(t-1) \tag{15}$$

where $S(t)$ is the cumulative demand in Eq. (13). If the system lifetime is exponential with failure rate α, and the fleet size increases in a Poisson counting process with rate λ, the mean and variance of $D(t)$ can be explicitly derived as follows [48],

$$E[D(t)] = \alpha(1+\lambda t)l + \frac{1}{2}\alpha\lambda l^2, \quad \text{for } 0 \le l \le t. \tag{16}$$

$$Var(D(t)) = \alpha(1+\lambda t)l + \left(\frac{1}{2}\alpha\lambda + a^2\lambda t\right)l^2 + \frac{1}{3}a^2\lambda l^3, \quad \text{for } 0 \le l \le t. \tag{17}$$

Below two cases are used to illustrate how Eqs. (5) and (6) can effectively capture the non-stationary behavior of $D(t)$. In Case 1, we have $\alpha = 0.1$ failure/week, $\lambda = 1$ system/week, and $l = 1$ week. In Case 2, we have $\alpha = 0.1$ failure/week, $\lambda = 0.5$ system/week, and $l = 3$ weeks. Figure 9 shows the mean and standard deviation (Stdev) of both cases. Two observations can be made. First, as expected, the mean demand and its standard deviation increase with the fleet size. For given α and λ, $E[D(t)]$ increases linearly with t, and the Stdv($D(t)$) goes up with $t^{1/2}$. Second, the value of l has a major impact on the safety stock level. For instance, the value of l in Case 2 is two times longer than that in Case 1. As a result, larger $E[D(t)]$ and $Stdv(D(t))$ are observed in Case 2 even if the fleet size of Case 2 is only half of Case 1.

Let $\overline{d}$ be the average spare parts demand per unit time during $[t - l, t]$, and can be estimated by

$$\overline{d} = \frac{E[D(t)]}{l} = \alpha(1+\lambda t) + \frac{1}{2}\alpha\lambda l = \alpha + \alpha\lambda(t + 0.5l) \tag{18}$$

As the inventory evolves, it implies $\overline{d}$ is independent of l because of $t + 0.5\ l \approx t$ for large t. Two necessary conditions are associated with an HPP: (1) the demand rate is constant; and (2) the variance-to-mean ratio (VMR) is unity. The first condition is stratified for $t >> l$. Below we show that the second condition also holds by computing the VMR of $D(t)$ as follows

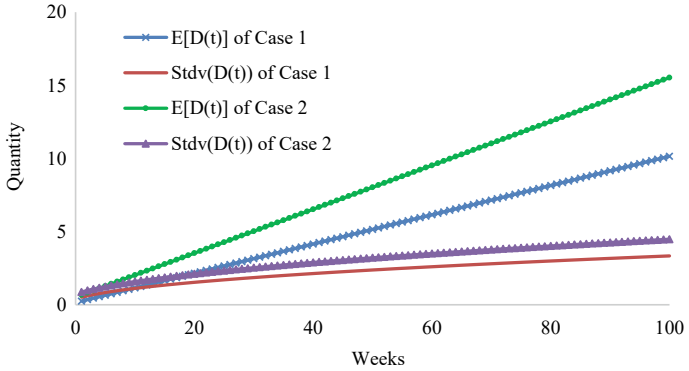


Fig. 9 Mean and standard deviation of lead time spares demands

$$VMR = \frac{Var(D(t))}{E[D(t)]} = 1 + \frac{\alpha^2\lambda t l^2 + \frac{1}{3}\alpha^2\lambda l^3}{\alpha l + \alpha\lambda t l + \frac{1}{2}\alpha\lambda l^2} < 1 + \frac{\alpha^2\lambda t l^2 + \frac{1}{3}\alpha^2\lambda l^3}{\alpha\lambda t l + \frac{1}{2}\alpha\lambda l^2} < 1 + \alpha l. \tag{19}$$

Equation (19) shows that VMR $\cong$ 1 if $\alpha l << 1$. For critical parts used in capital equipment, condition $\alpha l << 1$ generally holds because system's mean-time-to-failure ($1/\alpha$) is much larger than l, hence $\alpha l \approx 0$. Thus it can be concluded that the inventory lead time demand can be approximated as an HPP provided $\alpha l << 1$.

6 Lead Time Demand Analysis Using Simulation

In this section, we use Matlab software to develop a simulation program to characterize the spare parts demand under a growing fleet size. The purpose of the simulation is to verify whether the spares demand during the inventory replenishment lead time is a Poisson process. The key parameters in the simulation are the part replacement or failure rate α, new system installation rate λ, new product introduction period v, and inventory lead time l.

Table 3 shows the input data for the simulation, and the values are estimated based on the maintenance interval of Boeing 787 engines. For instance, if an engine is overhauled every 5000 flight hours, we have $\alpha = 365 \times 16/5000 = 1.168$ replacement/year assuming engine's daily operating time is 16 h. If engine's mean-time-to-overhaul is 10,000 h, the replacement rate α is 0.584. From 2011 to 2022, Boeing delivered a total of 1008 planes with 82 planes per year on average. Thus, we set λ to be 50, 100, and 200, respectively. Values are also appropriately assigned to v and l. The number of different scenarios or cases in simulation reaches $2 \times 3 \times 3 \times 2 = 36$.

Given a particular case, say $\alpha = 0.584$, $\lambda = 100$, $v = 5$, and $l = 2$, the simulation is repeatedly run 5000 times. The total number of runs for 36 cases is 5000 $\times$ 36

Table 3 Simulation parameters based on Boeing 787 fleet engines

Parameters	Values in simulation	Unit
α	{0.584, 1.1680}	Replacement/year
λ	{50, 100, 200}	Installation/year
v	{5, 10, 20}	Year
l	{2, 4}	Week

= 180,000. For each run, the program randomly generates plane installation time, engine replacement time, and spares demands for the planes installed between [0, v]. At the fleet level, the inter-arrival time between two consecutive replacements is recorded, and the lead time spares demands are tested against the HPP hypothesis. The confidence level of an HPP is estimated for each case and the result is shown in Table 4. Among the 36 cases, the lowest confidence level is found in Cases 7 and 34 with 94.9%, and the highest confidence level is observed in Cases 4 and 31 with 95.1%. The average confidence level is 95%. As a result, it can be stated that the lead time of spares demand of 36 cases follows the Poisson process with 95% confidence.

Analysis of Variance (ANOVA) is performed to examine how parameters α, λ, v, and l influence the HPP hypothesis test on the lead time demand in Table 5. Note that DF represents the degrees of freedom, SS stands for the sum of square, and MS is the mean of square. Since the P-value of all the parameters is larger than 0.05, indicating α, λ, v, and l are significant to the confidence level. Particularly, λ and l are the most influential to the lead time demand behavior because their P-values are much higher than those of α and v.

7 System Availability with Reliability-Maintenance-Inventory Integration

7.1 Single-Item System Availability

In previous sections, the repairable inventory system is characterized by two parallel Erlang-C queues: one handles failure returns and the other rejuvenates deteriorated items. The lead time spares demands are investigated under a fixed installed base and a growing installed base, respectively. In both cases, the Poisson demand behaviors are examined analytically and further verified by simulation. This section presents a system availability model that seamlessly integrates reliability-redundancy, maintenance time, spares inventory, repair and renewing capacity, and installed base information. We first estimate the single-item system availability, and then extend it to the multi-item system. A generic system availability model is given as follows

$$A = \frac{T_{MTBR}}{T_{MTBR} + T_{MDT}}, \tag{20}$$

Table 4 Simulation inputs and results for 36 cases

Case	α	λ	l	v	Confidence for HPP (%)
1	0.584	50	0.037	5	94.97
2	0.584	50	0.037	10	95.02
3	0.584	50	0.037	20	95.05
4	0.584	50	0.07407	5	95.11
5	0.584	50	0.07407	10	94.91
6	0.584	50	0.07407	20	95.03
7	0.584	100	0.037	5	94.89
8	0.584	100	0.037	10	94.99
9	0.584	100	0.037	20	94.97
10	0.584	100	0.07407	5	94.92
11	0.584	100	0.07407	10	94.99
12	0.584	100	0.07407	20	95.06
13	0.584	200	0.037	5	95.00
14	0.584	200	0.037	10	94.97
15	0.584	200	0.037	20	95.00
16	0.584	200	0.07407	5	94.95
17	0.584	200	0.07407	10	95.01
18	0.584	200	0.07407	20	94.99
19	1.168	50	0.037	5	94.92
20	1.168	50	0.037	10	95.04
21	1.168	50	0.037	20	94.95
22	1.168	50	0.07407	5	94.99
23	1.168	50	0.07407	10	95.01
24	1.168	50	0.07407	20	94.98
25	1.168	100	0.037	5	94.98
26	1.168	100	0.037	10	95.06
27	1.168	100	0.037	20	94.99
28	1.168	100	0.07407	5	95.01
29	1.168	100	0.07407	10	95.07
30	1.168	100	0.07407	20	95.01
31	1.168	200	0.037	5	95.13
32	1.168	200	0.037	10	94.99
33	1.168	200	0.037	20	95.05
34	1.168	200	0.07407	5	94.91
35	1.168	200	0.07407	10	95.03
36	1.168	200	0.07407	20	95.04

Table 5 ANOVA for analyzing Poisson lead time demand

Source	DF	SS	MS	*F*-value	*P*-value
α	1	0.00344	0.00344	1.1	0.303
λ	2	0.00113	0.00056	0.18	0.836
l	1	0.00010	0.00010	0.03	0.860
v	2	0.00610	0.00305	0.98	0.388
Error	29	0.09056	0.00312		
Total	35	0.10133			

where T_{MTBR} represents the mean time between replacements, and T_{MDT} represents the system mean downtime. Under age-based PM, the T_{MTBR} can be estimated by

$$T_{MTBR} = \int_0^\tau R(t)dt = \int_0^\tau (1 - F(t))dt = \tau - \int_0^\tau F(t)dt \tag{21}$$

If corrective maintenance policy is adopted, T_{MTBR} is equal to the mean-time-to-failure. To estimate T_{MDT}, both failure replacement and planned replacement shall be considered under the PM policy. Let T_q and T_p be system's failure downtime and planned downtime, respectively. Then,

$$T_q = t_s + t_q \Pr\{D > s\} \tag{22}$$

$$T_p = t_s + t_p \Pr\{D > s\} \tag{23}$$

where t_s is the hands-on replacement time, D is the lead time demand of the spares inventory, and s is the spares stock level. In Eq. (22), for example, T_q is equals to t_s if $D \leq$ s. When the spare is out of stock (i.e., $D >$ s), the downtime is prolonged because of waiting for the spare item. The average waiting time equals t_q given in Eq. (2). A similar interpretation can be applied to T_p and t_p is given in Eq. (4). For non-Markovian queues, Eqs. (5) and (6) shall be used to substitute for t_q and t_p. Now the system downtime can be further expressed as follows [49]

$$\begin{aligned} T_{MDT} &= T_q F(\tau) + T_p R(\tau) \\ &= \big(t_s + t_q \Pr\{D > s\}\big) F(\tau) + \big(t_s + t_p \Pr\{D > s\}\big) R(\tau) \\ &= t_s + \big(t_p R(\tau) + t_q F(\tau)\big) \Pr\{D > s\} \\ &= t_s + \left(\left(\frac{B_p}{p\mu_p - \lambda_p} + \frac{1}{\mu_p} \right) R(\tau) + \left(\frac{B_q}{q\mu_q - \lambda_q} + \frac{1}{\mu_q} \right) F(\tau) \right) \Pr\{D > s\} \end{aligned} \tag{24}$$

Note that $F(\tau)$ and $R(\tau)$ are the probability of failure and planned replacement at τ, respectively. By substituting Eqs. (21) and (24) into (20), the availability of a

single-item system can be estimated by [49]

$$A = \frac{\int_0^\tau R(t)dt}{\int_0^\tau R(t)dt + t_s + \left(\left(\frac{B_p}{p\mu_p - \lambda_p} + \frac{1}{\mu_p}\right)R(\tau) + \left(\frac{B_q}{q\mu_q - \lambda_q} + \frac{1}{\mu_q}\right)F(\tau)\right)\Pr\{D > s\}} \tag{25}$$

Equation (25) represents an integrated reliability-maintenance-inventory approach to modeling system availability. This is because reliability $R(t)$, maintenance time τ, spares inventory s, repair capacity q, and renewing capacity p are incorporated into a single formula. Note that B_q and B_p are given in Eqs. (1) and (3), respectively. In addition, there is no specific assumption on the inventory lead time demand.

7.2 Redundant System Availability

A redundant system is a type of multi-item system with the configuration of several identical items. For a k-out-n active redundant system, the system is available if k items are good for $k \leq n$. The availability of an active redundant system, denoted as A_{rd}, can be expressed as follows

$$A_{rd} = \sum_{i=k}^{n} \binom{n}{i} A^i (1 - A)^{n-i} \tag{26}$$

where A is the availability for the single-item system given in Eq. (25). The availability of a cold-standby system can also be estimated using Eq. (26). Unlike active redundant components, an item in cold standby is not in the operational state. If an active component fails, the cold-standby item is switched to the operational mode for the failure substitution. In addition, the spares demand rate of a cold-standby redundant system fleet is lower than that of active redundant systems. This is because the former has a smaller number of active items in operation at a given instant of time.

7.3 Numerical Experiment

We demonstrate the application of the integrated reliability-maintenance-inventory approach in an ATE system fleet. An ATE can be treated as a k-out-of-n active redundant system where k is the minimum required working items, and n is the total items in a system. Tables 6 shows the key parameters associated with design, manufacturing, maintenance, repair, and spares stocking for a type of ATE system. To meet the system availability target of 0.99, the OEM minimizes the annualized system

Table 6 Parameters for integrated reliability-maintenance-inventory optimization model

Parameters	Value	Unit	Comments
α	0.4	Failures/year	Weibull scale parameter
β	3.5	n/a	Weibull shape parameter
k	7	n/a	Minimum required working items
n	10	n/a	Total items in a system
r	0.05	Discount rate	Interest rate
t_s	24	Hour	Hands-on replacement time
t_p	9	Day	Renewing TAT
t_q	18	Day	Repairing TAT
c_{LRU}	100,000	$/item	Part or item unit cost
c_u	6000	$/renewing	Renewing cost per item
c_v	9000	$/repair	Repair cost per item
c_h	20,000	$/item/year	Inventory holding cost
c_p	480,000	$/server	Cost of renewing server
c_q	640,000	$/server	Cost of repairing sever

cost by optimizing redundant items x, maintenance time τ, spare parts inventory s, renewing servers p, and repair servers q under different sizes of install base. The detailed formulation of the reliability-maintenance-inventory optimization model can be referred to [49]. Figure 10 depicts the optimal solutions of x, τ, s, p, and q as the installed base m increases from 5 to 100.

Three observations are made from this numerical experiment. First, when the installed base is small, such as $m = 5$ or 20, component or item redundancy (i.e., $x = 1$) is preferred over spares inventory as it results in a lower annualized system cost. The annualized system cost includes: (1) the system's initial capital; (2) the overhead

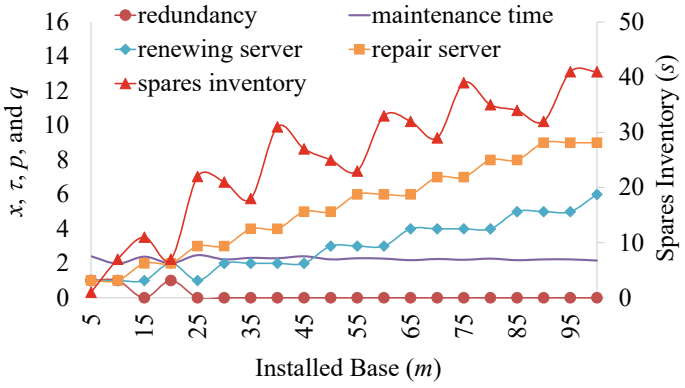


Fig. 10 Joint decision on redundancy, maintenance, spares, and repairs

of repairing and renewing parts; (3) the spare parts inventory; and (4) the expense of operating repair and renewal facilities. However, as m becomes larger, it is more economical to invest spares inventory, and repair and renewing capacities to achieve the system availability target. Second, maintenance time τ is relatively independent to the installed base. It is found that τ varies between 2 and 2.5 regardless of m. Third, the value of s, p, and q increases with m. This is expected because a larger installed base will generate more failure and planned replacements, hence requiring more spare parts, repair, and renewal services.

8 Conclusion

This study investigates the availability management of repairable systems across the design, manufacturing and after-sales market. The proposed system availability model seamlessly incorporates reliability, redundancy, maintenance, repair, spares inventory, and installed base into a unified formula. The study represents a first-of-its-kind in managing system availability through an integrated reliability-maintenance-inventory allocation approach. The method is applicable to a stationery system fleet or a growing installed base. The research findings can be highlighted in three aspects. First, the aggregate fleet spares demand under preventive maintenance can be decomposed into a failure stream and a planned replacement stream, each being approximated as the Poisson process. Second, when the installed base changes and increases, both the analytical model and the simulation results show that the inventory lead time demand can be treated as a Poisson process with 95 percent of confidence on average. Third, the numerical experiment shows that redundancy strategy is economical for a small installed base, while spares inventory is more cost effective for a large fleet size. As future efforts, we would like to expand the product-service supply chain network that involves multi-echelon, multi-location with parts transshipment. Another direction is to compare the product lifetime cost between reliability-maintenance-inventory integration and the separated management strategies.

References

1. Beuren FH, Ferreira MGG, Miguel PAC (2013) Product-service systems: a literature review on integrated products and services. J Clean Prod 47:222–231
2. Zhang J, Zhao X, Song Y, Qiu Q (2021) Joint optimization of maintenance and spares ordering policy for a use-oriented product-service system with multiple failure modes. Appl Stoch Model Bus Ind 37(6):1123–1142
3. Zhang L, Feng L, Wang J, Lin KY (2022) Integration of design, manufacturing, and service based on digital twin to realize intelligent manufacturing. Machines 10(4):275
4. Chern MS (1992) On the computational complexity of reliability redundancy allocation in a series system. Oper Res Lett 11(5):309–315

5. Kuo W, Wan R (2007) Recent advances in optimal reliability allocation. IEEE Trans Syst Man Cybern Part A Syst Hum 37(2):143–156
6. Coit DW, Zio E (2019) The evolution of system reliability optimization. Reliab Eng Syst Saf 192:106259
7. Si S, Zhao J, Cai Z, Dui H (2020) Recent advances in system reliability optimization driven by importance measures. Front Eng Manage 7(3):335–358
8. Kuo W, Lin HH, Xu Z, Zhang W (1987) Reliability optimization with the Lagrange-multiplier and branch-and-bound technique. IEEE Trans Reliab 36(5):624–630
9. Coit DW, Smith AE (1996) Reliability optimization of series-parallel systems using a genetic algorithm. IEEE Trans Reliab 45(2):254–260
10. Kulturel-Konak S, Smith AE, Coit DW (2003) Efficiently solving the redundancy allocation problem using tabu search. IIE Trans 35(6):515–526
11. Yeh WC, Hsieh TJ (2011) Solving reliability redundancy allocation problems using an artificial bee colony algorithm. Comput Oper Res 38(11):1465–1473
12. Coit DW, Jin T, Wattanapongsakorn N (2004) System optimization with component reliability estimation uncertainty: a multi-criteria approach. IEEE Trans Reliab 53(3):369–380
13. Ardakan MA, Rezvan MT (2018) Multi-objective optimization of reliability–redundancy allocation problem with cold-standby strategy using NSGA-II. Reliab Eng Syst Saf 172:225–238
14. Si S, Liu M, Jiang Z, Jin T, Cai Z (2019) System reliability allocation and optimization based on generalized Birnbaum importance measure. IEEE Trans Reliab 68(3):831–843
15. Dursun İ, Akçay A, Van Houtum GJ (2022) Age-based maintenance under population heterogeneity: optimal exploration and exploitation. Eur J Oper Res 301(3):1007–1020
16. Jin L, Yamamoto W (2017) Adaptive age replacement using on-line monitoring. Procedia Eng 174:117–125
17. Zhao X, Al-Khalifa KN, Hamouda AM, Nakagawa T (2017) Age replacement models: a summary with new perspectives and methods. Reliab Eng Syst Saf 161:95–105
18. Jiang R (2019) Risk-sensitive cost models of age replacement policy: comments and proposals. J Oper Res Soc 70(4):548–554
19. Peng Y, Dong M, Zuo MJ (2010) Current status of machine prognostics in condition-based maintenance: a review. Int J Adv Manuf Technol 50(1):297–313
20. Zhao Q, Jia X, Cheng ZJ, Guo B (2018) Bayes estimation of residual life by fusing multisource information. Front Eng Manage 5(4):524–532
21. Zhu W, Fouladirad M, Bérenguer C (2016) A multi-level maintenance policy for a multi-component and multifailure mode system with two independent failure modes. Reliab Eng Syst Saf 153:50–63
22. Alaswad S, Xiang Y (2017) A review on condition-based maintenance optimization models for stochastically deteriorating system. Reliab Eng Syst Saf 157:54–63
23. Li Y, Peng S, Li Y, Jiang W (2020) A review of condition-based maintenance: its prognostic and operational aspects. Front Eng Manage 7(3):323–334
24. Hu Y, Miao X, Si Y, Pan E, Zio E (2022) Prognostics and health management: a review from the perspectives of design, development and decision. Reliab Eng Syst Saf 217:108063
25. Sherbrooke CC (1968) METRIC: A multi-echelon technique for recoverable item control. Oper Res 16(1):122–141
26. Muckstadt JA (1973) A model for a multi-item, multi-echelon, multi-indenture inventory system. Manage Sci 20(4-part-i):472–481
27. Lee HL (1987) A multi-echelon inventory model for repairable items with emergency lateral transshipments. Manage Sci 33(10):1302–1316
28. Alfredsson P (1997) Optimization of multi-echelon repairable item inventory systems with simultaneous location of repair facilities. Eur J Oper Res 99(3):584–595
29. Lau HC, Song H (2008) Multi-echelon repairable item inventory system with limited repair capacity under nonstationary demands. Int J Inventory Res 1(1):67
30. de Smidt-Destombes KS, van der Heijden MC, van Harten A (2009) Joint optimisation of spare part inventory, maintenance frequency and repair capacity for k-out-of-N systems. Int J Prod Econ 118(1):260–268

31. Jin T, Tian Z, Xie M (2015) A game-theoretical approach for optimizing maintenance, spares and service capacity in performance contracting. Int J Prod Econ 161:31–43
32. Basten RJ, Ryan JK (2019) The value of maintenance delay flexibility for improved spare parts inventory management. Eur J Oper Res 278(2):646–657
33. Zhu W, Castanier B, Bettayeb B (2019) A dynamic programming-based maintenance model of offshore wind turbine considering logistic delay and weather condition. Reliab Eng Syst Saf 190:106512
34. Basten RJI, van Houtum GJ (2014) System-oriented inventory models for spare parts. Surv Oper Res Manage Sci 19(1):34–55
35. Mouschoutzi M, Ponis ST (2022) A comprehensive literature review on spare parts logistics management in the maritime industry. Asian J Shipping Logistics
36. Jin T (2019) Reliability engineering and services. Wiley, New Jersey Chapter 10
37. Diaz A, Fu MC (1997) Models for multi-echelon repairable item inventory systems with limited repair capacity. Eur J Oper Res 97(3):480–492
38. Winston W (2004) Operations research: applications and algorithms. 4th edn. Brooke/Cole Cengage Learning, Belmont (Chapter 20)
39. Allen AO (1978) Probability, statistics and queueing theory, with computer science applications. Academic Press, New York
40. Cox DR, Smith WL (1954) On the superposition of renewal processes. Biometrika 41:91–99
41. Wang W (2012) A stochastic model for joint spare parts inventory and planned maintenance optimisation. Eur J Oper Res 216(1):127–139
42. Southwest Airlines (2022) Southwest airlines fleet details and history, available at https://www.planespotters.net/airline/Southwest-Airlines. Accessed on 8 Nov 2022
43. Wu S (2019) Superimposed renewal processes in reliability. Wiley Stats Ref: Statistics Reference https://doi.org/10.1002/9781118445112.stat08228
44. Jin T, Li H, Sun F (2021) System availability considering redundancy, maintenance and spare parts with dual repair processes. In: Proceedings of the IIE annual conference. Institute of Industrial and Systems Engineers (IISE), pp 590–595
45. Wikipedia (2022) Boeing 787 Dreamliner, available at https://en.wikipedia.org/wiki/Boeing_787_Dreamliner. Accessed on 2 Oct 2022
46. Scott A (2016) Japanese airline ANA to replace 100 rolls engines on 787s. Aerospace and defense, August 31, 2016, available at https://www.reuters.com. Last Accessed on 4 Oct 2022
47. Jin T, Liao H (2009) Spare parts inventory control considering stochastic growth of an installed base. Comput Ind Eng 56(1):452–460
48. Jin T, Taboada H, Espiritu J, Liao H (2017) Allocation of reliability–redundancy and spares inventory under Poisson fleet expansion. IISE Trans 49(7):737–751
49. Jin T, Zhu W (2022) Optimizing redundancy, maintenance and spare parts for system availability under decentralized repair. Texas State University Working Paper

Use of Artificial Neural Networks to Enhance Container Port Safety Analysis Under Uncertainty

Hani Al Yami, Ramin Riahi, Jin Wang, and Zaili Yang

Abstract This chapter proposes a modified failure mode effect analysis (FMEA) approach using Artificial Neural Networks (ANNs) to evaluate and predict the operational risks of container terminals. It effectively integrates two established methods in one framework to realise complex risk analysis from a whole system perspective, including fuzzy rule based Bayesian networks (FRBN) for risk analysis of particular hazards in ports and fuzzy evidential reasoning (FER) for safety evaluation of ports in a systematic way. During this process, ANNs are integrated with FRBN and FER respectively to create two sub-models. The first sub-model is FRBN-ANN that incorporates Bayesian networks (BNs) with ANNs to facilitate risk prediction of each identified hazard in a container port. The second sub-model is FER-ANN, which uses ANNs to simulate the FER method to ease the aggregation of all the hazards to obtain the safety level of the port. Finally, the two sub-models are combined into a single safety model, which can help simplify risk prediction, and realise real-time safety evaluation of ports at hazard or whole system levels. The Levenberg–Marquardt (trainlm) back-propagation algorithm trial and error approach was used to determine the optimal ANN architecture. The proposed ANN model produced small deviations that indicate high predictive accuracy with satisfactory determination coefficients (i.e., the regression) for forecasting operational risks of container ports. It provides an effective risk prediction tool for complex port safety systems, and significantly simplifies the port safety analysis and prediction in a feasible, versatile, and accurate manner. It, through the black box approach of ANN, provides a mathematically unsophisticated solution and hence aids the visualisation of risk analysis outcomes without the need of the end users to understand the complicated computing process

H. Al Yami
Faculty of Maritime Studies, King Abdulaziz University, Jeddah, Saudi Arabia

R. Riahi
Columbia Shipping Management (Deutschland) GmbH, Hamburg, Germany

J. Wang · Z. Yang (✉)
Liverpool Logistics Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, UK
e-mail: z.yang@ljmu.ac.uk

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_11

of the risk inference. It makes significant contributions to port safety analysis and management in practice.

1 Introduction

Innovative technologies, advanced safety management methodologies, and hazards-driven risk mitigation regimes have resulted in increasing safety standards in container ports. Advanced risk models such as fuzzy rule-based Bayesian networks (FRBN) [1] and fuzzy evidential reasoning (FER) [2] focusing much on result accuracy and reliability assurance, often compromise their easiness, hence revealing problems in their practical applications. New approaches that enable to simplify the complicated risk analysis and prediction process become necessary. Artificial Neural Networks (ANNs) have been successfully applied for real-time risk prediction in various sectors over the past decade, due to their reliable, robust, and salient capturing of non-linear relationships between complex system variables (i.e., multi-input/output). On the other hand, failure mode and effect analysis (FMEA) is one of the most widely applied hazard identification and risk analysis methods, due to its visibility and easiness. ANNs can enhance FMEA performance by overcoming its incapability of tackling data uncertainty; at the same time, it can ease stakeholders' burden of handling a complex large amount of data to measure, predict, and improve system safety and reliability performance. Research has shown that ANNs have powerful pattern classification and recognition capabilities. Inspired by biological systems, particularly research into the human brain as a large-scale nonlinear drive system, ANNs offer a computational paradigm that learns and generalises from experience. It also has many egregious functions, such as adaptive learning, real time operation, self-organisation, thinking and reasoning, judging and memory, and fault tolerance [3, 4]. Since the 1980s, research on ANNs has made remarkable developments, and has been successfully applied in a wide variety of domains, including risk and safety studies, but not yet in port risk and safety studies.

In the past decade, applications of ANNs in risk assessment are seen in medical research [5–7], financial investigations [6, 8–12], and civil engineering studies [13–18]. However, very few studies use ANN in risk analysis in maritime systems. Ung et al. [19] applied ANNs to predict the risk level of sea-lane navigation within port areas by incorporating fuzzy set theory and ANNs. Although, showing a unique conception of applying ANNs in the maritime port industry, however, the approach has only considered the fuzziness, incapable of modelling the other types of uncertainties in data (e.g. incompleteness and randomness), which releases a research gap.

Although FRBN and FER have proven their ability to evaluate risks in container ports, their complexity in handling a large amount of data and in dealing with two separate software packages forces the stakeholders to navigate non-user-friendly processes and reduces their applicability in practice. This chapter aims to develop an integrated risk analysis method for container ports, using ANNs to predict and

evaluate the criticality of hazardous events (HEs) in a container port. The ANN approach is used to simplify and integrate two established methods in one framework to realise complex risk analysis from a whole system perspective. They include FRBN for risk analysis of particular hazards in ports [20] and FER for safety evaluation of ports from a systematic way [21]. Furthermore, we implement experimental data to train ANNs in a rational structure and develop an applicable, new risk-based decision support tool for risk prediction of container ports in practice.

In order to clearly map and explain the proposed models, this chapter is organised as follows. Section 2 introduces ANN principles and the relevant background information. Section 3 describes the methodology of a novel ANN framework. Section 4 uses ANN to simulate FRBN so as to evaluate the criticality of each identified HE in a container port locally. Section 5 uses ANN to simulate FER so as to evaluate the safety of the port system as well as the HEs of the most risk contribution to the system safety level globally. Section 6 presents the integrated ANN model to predict the risk index of each HE and provide an overall safety evaluation of the operational systems of a container port. Section 7 concludes the chapter with the possible future studies.

2 Artificial Neural Networks (ANNs)

In 1942, McCulloch and Pitts [22] proposed modelling neural nets as a single neuron form in terms of the computational "nervous activity" model, which describes the neuron as a linear threshold-computing unit with multiple inputs and a single output to solve character recognition problems [23]. In 1949, Hebb built the missing link between single neurons and network in his classic book *The Organization of Behaviour*. Rosenblatt developed a network in 1958 using McCulloch and Pitt's model, based on a unit called the "perceptron" [24]. Rosenblatt [25] and others explored and developed many types of perceptron based ANNs in the 1960s. The topic rapidly faded in the 1970s, however, because of two main problems: first, the practical difficulties of solving many real-world problems; and secondly, the serious limitations among perceptron which could not be solved by simply adding neuron layers [26]. It was also determined that the perceptron was incapable of representing simple, linearly inseparable functions, as in the famous "exclusive or" (XOR) problem [26, 27]. However, the primary problem was the absence of any learning algorithm to train such networks.

Hopfield [28] poured new life into this field by introducing two key concepts that overcame Minsky and Papert's identified limitations: first, the nonlinearity between total input received by a neuron and its produced output; secondly, the possibility of feedback coupling outputs with inputs [27]. Since then, ANNs have seen an explosion of interest, together with a paradigm change in recent years. They were intensively and extensively used as problem solving algorithms for application development, rather than accurate representations of the human nervous system [27, 29]. They have been successfully applied across an extraordinary range of domains [30–35].

A review by Liao and Wen [29], based 10,120 articles about ANN methodologies and application developments from 1995 to 2005, uses data mining to disclose the wide range of ANN applications in many fields of studies. Other researchers have explored the use of hybrid ANNs with deferent methods, such as neuro-fuzzy for time series modelling [36], neuro-fuzzy rule-based for stock market decision support modelling [37], Bayesian neural networks for medicine [38], and Dempster-Shafer neural network for navigation technology [39]. This sweeping success are largely attributed to its advantages, explored in Siegel et al. [40], Haykin [41] and Taha [42].

ANNs are computational modelling tools with flexible structures that capture and simulate complex input/output relationships. They are comprised of densely interconnected adaptive and simple processing elements, capable of performing massive parallel computations for data processing and knowledge representation [29, 43, 44]. The ANN terminology has been developed from a biological model that uses artificial neurons to imitate the learning process of the human brain (i.e., natural neurons) to a system that processes nonlinear and complex data, even when the data are imprecise and noisy. However, solving complex problems requires knowledge of biological network functionality rather than a replication of biological system operation [23, 29].

The complexity of real neurons is highly abstracted when modelling artificial neurons. ANNs consist of inputs (synapses), which are multiplied by weights (strength of respective signals), and then computed by a mathematical function, which determines the activation of the neuron; then another function (possibly identity) computes the artificial neuron's output. As a result, ANNs' mechanisms combine all artificial neurons to process information. The greater an artificial neuron's weight is, the stronger its input. A neuron's computation depends on weights and differs if the weight changes, because the weight is multiplied by input. By adjusting an artificial neuron's weights, the output can be obtained as desired for specific inputs. However, when an ANN consists of hundreds or thousands of neurons, manually finding all the necessary weights becomes complicated. Weight algorithms can find and adjust ANN weights in order to obtain desired network output. This process of weight adjustment is called learning or training [41].

A complex system may be deconstructed into simpler elements in order to understand and handle it. Simple elements can then be gathered to produce a complex system; networks represent one approach for achieving this [45]. There are many network types, but they all contain the following components:

- A set of nodes: nodes can be seen as computational units. They receive and process inputs to obtain an output. This processing may be very simple (summing the puts) or quite complex (a node might contain another network).
- Connections between nodes: connections determine information flow between nodes and can be unidirectional, when information flows only in one sense, and bidirectional, when information flows in either sense.

Node interactions through connections lead to a network's global behaviour, which cannot be observed through the network's elements. This global behaviour

is described as "emergent," meaning that the networks abilities supersede those of its elements, making networks a very powerful tool [41].

A neuron is a real function of the input vector $(x_1 \ldots x_j)$. The output y is obtained as

$$f(y_j) = f\left(\alpha + \sum_{i=1}^{k} w_{kj} x_j\right) \tag{1}$$

where f is a function (functions will be explained in detail in Sect. 3.1), x_1, x_2, x_3,….. x_j are the input signals, w_{k1}, w_{k2}, w_{k3}, …w_{kj} are the synaptic weights of the *kth* neuron, and α is the bias.

A graphical presentation of a neuron is given in Fig. 1. Mathematically, a multi-layer perceptron network is a function consisting of compositions of functions' weighted sums corresponding to neurons [41].

ANNs as a data processing system consist of a large number of simple, highly interconnected processing elements in an architecture inspired by the brain's cerebral cortex structure, and there are several architecture ANN types. Simpson [46] lists 26 different types of ANNs, Maren [47] lists 48, and Pham [48] estimates more than 50. Some networks are more proficient in solving perceptual problems, while others are more suitable for data modelling and functional approximation, but feed forward networks (e.g., Back-Propagation network) and recurrent networks are among the most widely used [23].

In Fig. 2, the BP information flows in one direction along connecting pathways, from the input layer via the hidden layers to the final output layer. There is no feedback (i.e., all links are unidirectional and there are no same layer neuron-to-neuron connections), and the output of any layer does not affect that same or preceding layer. These networks are the most widely used types and are considered the workhorse of

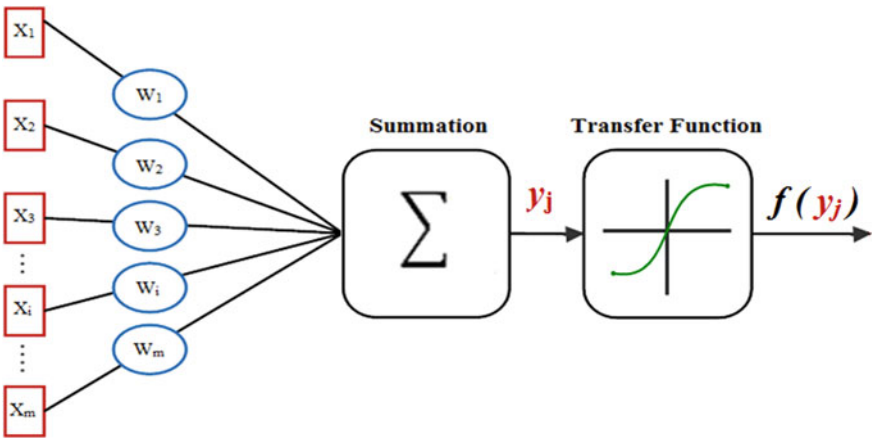


Fig. 1 A single neuron

ANNs because of their flexibility and adaptability in modelling a wide spectrum of problems in many application areas [49].

The recurrent network in Fig. 3 differs from feed forward network architectures in that there is at least one feedback loop. Thus, these networks have one layer with feedback connections; they may also have neurons with self-feedback links (i.e., a neuron's output is fed back into itself as input).

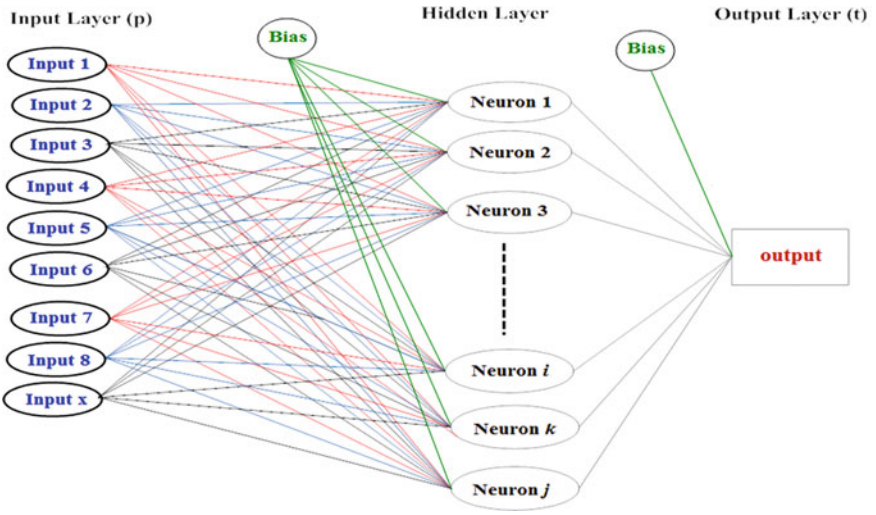


Fig. 2 Feed forward network

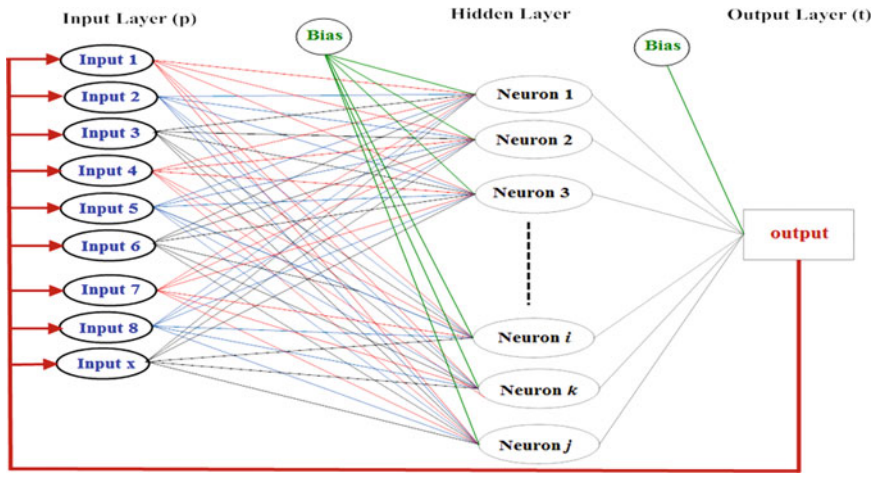


Fig. 3 Recurrent network

This study uses BP because of three main reasons: its ability to learn mapping from one data space to another using examples; high accuracy in capturing data's nonlinearity in (i.e., the relationship between inputs and outputs); and the simplicity in searching, accelerating, and stabilising the training process.

Multiple types of neural network software have been developed, including Environment for Computer Aided Neural Software Engineering (ECANSE) [50]; MATLAB: Neural Network Toolbar [51]; Neuroshell 1 and 2 [52]; and Statistica Neural Network [52]. Among ANN software, Matlab is a high-level language and interactive environment for numerical computation, visualisation, and programming that allows for analysing data, developing algorithms, and creating models and applications. Matlab computer language, tools, and built-in math functions can explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages. It also has the functions to integrate based algorithms with external applications and other programming languages. In this work, the Neural Network Toolbox Version 7.8 of MATLAB® mathematical software predicts risk evaluation for a container terminal.

3 Methodology

The purpose of this section is to propose the framework of using ANNs to combine the two separate methods of FRBN and FER, which were developed to evaluate container port risks at hazard and port system levels, respectively. The FRBN [20] provides a realistic and flexible method to describe input failure information for specific risk estimates of individual HEs at the bottom level of a risk analysis hierarchy as shown in Fig. 4, which contains 24 HEs in container ports.

It was found that when considering the risk levels of individual HEs, the most significant include [20]:

- Collision between the quay crane and the ship (HE4).
- Collision between two quay cranes (HE5).
- Crane break down due to human error (HE6).
- Person slips, trips and falls whilst working on surfaces with presence of oils (HE15).
- Collision between Terminal Tractor (TT) and trailer (HE1).
- Person slips, trips and falls whilst working on surfaces with presence of leaking cargo (HE13).
- Moving the crane without raising the Boom (lifting arm) of the gantry crane (HE7).

While the FER approach [21] is used to aggregate the risk estimates of all HEs collectively, allowing safety evaluation of container ports from a systematic perspective. As a result, the most significant HEs are as follows.

- Ignition sources from equipment near dangerous goods premises (HE9).

Fig. 4 Hierarchy of 24 significant hazards of container terminal operations [21]

- Leakage/ emission of dangerous goods from a container (HE8).
- Crane break down due to human error (HE6).
- Moving the crane without raising the boom (lifting arm) of the gantry crane (HE7).
- Person struck by falling object/s (HE16).
- Person handling dangerous goods in containers that have not been declared (HE17).

Although showing some attractiveness in port risk analysis, the two models have been criticised in terms of their complex inference processes. The proposed models using ANNs are constructed to ease the risk inference analysis of ports at both HE and system levels. During this process, HE6, HE7 and HE16 are selected for the demonstration purpose due to their high-risk indexes.

A multilayer perceptron neural network model is designed, consisting of an experimental data (ED) set collection used for model training and testing, network creation and configuration based on pre-processing and analysis of the data set, network training and validation, and finally, simulations and predictions [53]

The three parts for developing novel ANNs to model FRBN and FER are outlined in three steps as follows.

1. Design the ANN model to simulate the FRBN method in Alyami et al. [21]. FRBN has 12 inputs based on the four risk parameters (i.e., Likelihood, Consequence, Probability of hazards not being detected, and Impact of the hazard to the port system safety) and the linguistic terms for each risk parameter (i.e., High, Medium, and Low). The risk evaluation output for each identified HE has three linguistic terms (i.e., High, Medium and Low).
2. Design the ANN model to simulate the FER method. FER has nine inputs based on three HEs (i.e., HE6, HE7 and HE16) of high-risk evaluations resulting from the ANN in Step 1 and each of the investigated HE has three risk parameters (i.e., High, Medium, and Low). It has only one output processed in FER.
3. Construct the ANN model to simulate the combined the ANN models related to FRBN and FER in Steps 1 and 2 respectively. It creates a risk prediction tool that provides a panoramic view of the safety level of a container port's operation performance.

3.1 Algorithms of Modelling Performance Criteria

In a typical ANN, the input layer is composed of the ED (X_i), which is associated with the input layer's neurons (1, 2…,*i*,… *m*). The input signals are fed into the input layer, then transferred to the hidden layers' neurons (1, 2…, *j*,… *n*), where processing takes place by multiplying connection weights (w_{ij}) between two neurons and using the summation function to deliver output signals to the output layers (1, 2…, *k*,… *p*) [54].

Each layer's input data is processed to outputs using an activation (i.e., transfer) function, a nonlinear mathematical function known as a "transfer function." The most

widely used transfer functions are *tansig*, logarithmic sigmoid *(logsig)*, and *purelin*, described respectively below and illustrated in Fig. 5. The *tansig* activation function offers slightly better predictions than the others and is most commonly used in the hidden layer and the *purelin* activation function is used in the output layer [55–57].

Each layer's input data is processed to outputs using an activation (i.e., transfer) function, a nonlinear mathematical function known as a "transfer function". The most widely used transfer functions are *tansig*, logarithmic sigmoid *(logsig)*, and *purelin*, described respectively below and illustrated in Fig. 5. The *tansig* activation function offers slightly better predictions than the others and is most commonly used in the hidden layer and the *purelin* activation function is used in the output layer [55–57].

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{2}$$

$$f(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}} \tag{3}$$

$$f(x) = x \tag{4}$$

The data flow process for a single neuron in the network starts with each input streaming multiplied by a weight (w) and summed using the summation function. Then, this single value is processed through a transfer function to produce the output value of a neuron, as illustrated in Fig. 6 [55].

Selecting a training algorithm and activation function is crucial for robust ANN model performance. In general, linear functions are used for input and output layers, and nonlinear transfer functions for hidden layers [58].

Levenberg Marquardt Back Propagation (LMBP) is the most widely used optimisation algorithm for a variety of ANN problems [53, 58–62]. The LMBPA optimisation is a standard technique for nonlinear, least square problems, and was applied in this study to simulate actual risk estimation values for constructing a risk prediction system. The LMBPA data training algorithm procedure is as follows [23, 61, 63]

1. Initialise the weights and thresholds in the hidden and output layers, often in the range of $[-1, 1]$.
2. Calculate the hidden layer y value:

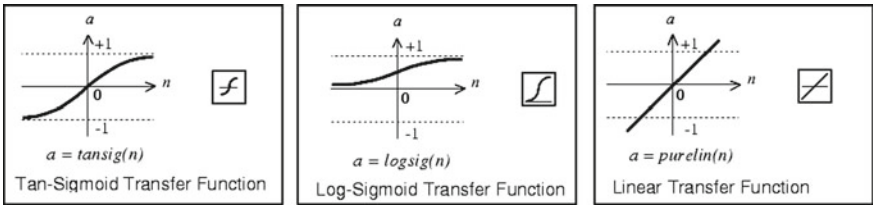


Fig. 5 Graphical representation for activation (transfer) functions

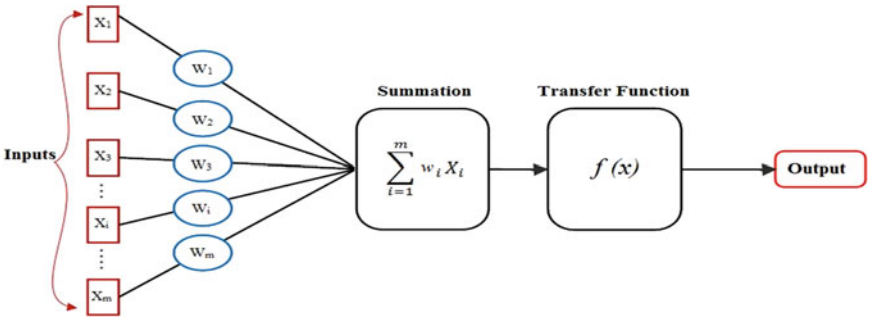


Fig. 6 Data flow process in a neuron

$$y_j(p) = \tan \text{ sig}\left[\sum_{i=1}^{n} x_i(p)w_{ij}(p) - \theta_j\right] \tag{5}$$

where *tansig* is the transfer function. n is the number of neurons in the hidden layer, x_{ij} is the i^{th} inputs, w_{ij} is the weight of x_{ij}, and θ is the threshold value.

3. Calculate the output layer y value:

$$y_k(p) = \text{tansig}\left[\sum_{j=1}^{n} x_{ik}(p)w_{jk}(p) - \theta_k\right] \tag{6}$$

where m is the number of neurons in the hidden layer and θ is the threshold value.

4. Calculate the output layer error:

$$\delta_k(p) = y_k(p)[1 - y_k(p)]e_k(p) \tag{7}$$

$$e_k(p) = y_{d,k}(p) - y_k(p) \tag{8}$$

5. Correct the output layer weight w:

$$\Delta w_{jk}(p) = \alpha y_j(p)\delta(p) \tag{9}$$

$$\Delta w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \tag{10}$$

6. Calculate the hidden layer error δ:

$$\delta(p) = y_j(p)\left[1 - y_j(p)\right]\sum_{k=1}^{1}\delta_k(p)w_{jk}(p) \tag{11}$$

7. Correct the hidden weight *w*:

$$\Delta w_{ij}(p) = \alpha x_i(p)\delta_j(p) \tag{12}$$

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \tag{13}$$

The steepest descendent method has a very fast convergence speed; however, when the optimal point is reached due to a decreasing gradient, the convergence speed slows. Therefore, the Newton method is integrated with the steepest descendent method to obtain excellent convergence effects when approaching the optimal point [64]. Accordingly, the performance function has the form of a square sum and represents the Hessian matrix, which is as follows:

$$H = J^T J \tag{14}$$

$$g = J^T e \tag{15}$$

where J is the Jacobian matrix, containing the first order differentiation of the network error against weight and partial weight, and

e is the network error vector and g is the gradient.

The basic principles of the Newton Method are:

$$X_{k+1} = X_k - A_k^{-1} g_k \tag{16}$$

where A_k is the Hessian matrix, namely, the second order differentiation of the performance function in the weights and partial weights.

$$A_k = \nabla^2 f(X) \tag{17}$$

$$g_k = \nabla f(X) \tag{18}$$

The LMBP algorithm uses the Hessian matrix value to correct the Newton method:

$$X_{k+1} = X_k - \left[J^T + \mu I\right]^{-1} J^T e \tag{19}$$

where parameter μ ensures that matrix inversion will always produce a result, and this parameter will depend on evaluation of sum of squared errors.

3.2 Algorithms of Modelling Assessment Criteria

Output values are compared with target values (i.e., experimental results) to assess model predictions. The differences between predicted and target values are evaluated against the modelling performance criteria established within the ANN algorithm. Hence, it is necessary to reprocess output values if modelling performance criteria are not met [54, 55].

The Minimum Mean Squared Errors (MSE) and the R^2 value are the most common performance criteria for the ANN model performance evaluation. The optimum number of neurons was determined by the minimum MSE value from the training and prediction dataset. The MSE represents the difference between an approximating function $F(w,x_i)$ of the adjustable weight (w) for the predicted values and target values (i.e., the error) with a range from 0 to 1 where the lower values of MSE are preferable [55, 59, 60]. R^2 shows the percentage of variability between ED and predicted data. R^2 values range between zero and one, in which R2 value ≈ 1 means a greater correlation and stronger relationship between predicted and actual values [55]. The MSE and R^2 values provide information on general error ranges between predicted and target values.

The above criteria are commonly used for validating models and their predictions. Notably, however, the ED quality is an essential requirement for modelling work; otherwise, the results of statistical tests and model predictions will be inaccurate [65].

3.3 Development of ANN Modelling in Maritime Ports

The FRBN approach evaluates the criticality of the 24 HEs in a container terminal, using four risk parameters: HE occurrence probability (L), HE consequences/severity (C), the probability of HE being undetected (P) and HE impacts on the resilience of port operational systems (I). The four risk parameters are constructed to form the IF part in an IF–THEN rule base, while the risk estimate (R) of hazards is presented in the THEN part associated with three linguistic grades (i.e., High, Medium and Low). Degree of Belief (DoB) of High (H), Medium (M), and Low (L) are employed to describe L, C, P, I, and R. The degrees of the parameters, calculated by the FRBN method for each HE, are based on knowledge accumulated from past events, taking into account domain experts' judgements. The process is illustrated in Fig. 7.

FERR uses the results from FRBN and the evidential reasoning method to aggregate the 24 HEs and evaluate their collective criticality in a container port. The risk level of the System (RoS) is a single value for a container port, shown in Fig. 8.

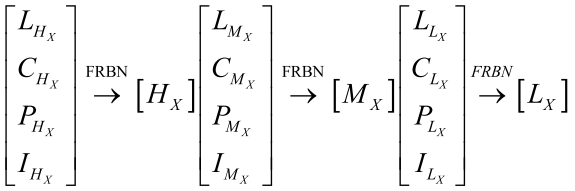


Fig. 7 FRBN process (where $(X = 1, 2 \ldots 24)$ is the HE number.)

$$\begin{bmatrix} L_{H_X} \\ C_{H_X} \\ P_{H_X} \\ I_{H_X} \end{bmatrix} \xrightarrow{\text{FRBN}} \begin{bmatrix} H_X \end{bmatrix} \begin{bmatrix} L_{M_X} \\ C_{M_X} \\ P_{M_X} \\ I_{M_X} \end{bmatrix} \xrightarrow{\text{FRBN}} \begin{bmatrix} M_X \end{bmatrix} \begin{bmatrix} L_{L_X} \\ C_{L_X} \\ P_{L_X} \\ I_{L_X} \end{bmatrix} \xrightarrow{FRBN} \begin{bmatrix} L_X \end{bmatrix}$$

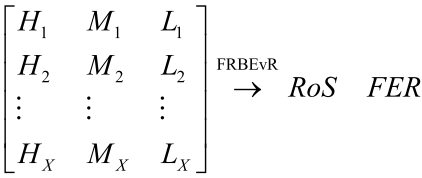


Fig. 8 FRBEvR process with ANNs

$$\begin{bmatrix} H_1 & M_1 & L_1 \\ H_2 & M_2 & L_2 \\ \vdots & \vdots & \vdots \\ H_X & M_X & L_X \end{bmatrix} \xrightarrow{\text{FRBEvR}} RoS \quad FER$$

Two three-layers ANN models were developed to predict risk evaluation of a container port operation for each HE individually and aggregated collectively. An optimal network architecture was determined to include one hidden layer with 40 neurons using the LMBP algorithm, with transfer function *tansig* at the hidden layer and transfer function *purelin* at the output layer.

Although the two models have the same features, they differ in trajectories concerning input and the output structures. The number of neurons in the input layer is twelve and nine, while the number of neurons in the output layer is three and one for the ANNs of FRBN and FER, respectively. As a result, the network architectures for are constructed as (12–40–3) and (9–40–1), respectively. It is noteworthy that there are no clear guidelines for choosing an appropriate number of neurons in the hidden layer; this is generally optimised by trial and error [53–58, 60, 65–68].

3.4 Experimental Data Processing

The EDs in this study are obtained (i.e., actual data for testing) from previous studies [20, 21]. As aforementioned, the relationship between the ANNs for the FRBN and FER approaches is that the outputs of the former are used as inputs for the latter.

As mentioned in the introduction of Sect. 3, HE6, HE7, and HE16 are selected for the development of the ANNs given their high-risk indexes from previous relevant studies. Although it was challenging to obtain the required EDs for training ANNs due to lack of objective failure data in container terminals, EDs were created and obtained using the Python program [69, 70]. The generated EDs' objective is to make the relationship between inputs and outputs maximally informative, while ensuring that the EDs adequately cover the region between zero and one with plausible inference intervals and maintaining each model's characteristics. Above all, the constructed models based on generated EDs makes them applicable for any

other maritime container terminal for risk evaluation and prediction. More details on generating EDs and the ANNs are presented in Sects. 4–6.

4 The ANN Model to Simulate FRBN for Risk Analysis of HE in Container Ports

The ANN model consists of 12 inputs representing the four risk parameters—L, C, P, and I, each of which has three grades (i.e., High, Medium and Low) and three outputs representing the R (risk estimate of the identified HE6, HE7 and HE16) in FRBN (i.e., ANN target). The simulation of FRBN using ANNs include the three steps of ED analysis, model optimisation and results validation.

4.1 ED Analysis of FRBN-ANN

To generate the ED to train and test the ANN model with the best prediction results, the simulated processes (i.e., predicted input and output values) should be inside the variables' domain, meaning that every possible risk parameter assessment should be included. Therefore, a 0.2 inference interval variation is applied to the DoB of each the three grades of each risk parameter, which not only narrows the range of deviation between input and consequently output values, but also adequately increases the training dataset.

Transferring the inference interval of 0.2 DoB from zero to one resulted in 21 possible combinations among the associated three grades of each of the four risk parameters. Accordingly, the sum of all possible combinations is calculated as 194,481 (=$21 \times 21 \times 21 \times 21$). As a result, the ED containing 194,481 sets, with 12 inputs and three outputs in each set, is obtained and partially shown in Table 1.

4.2 Optimisation of the FRBN-ANN Model

After selecting *trainlm* as the training algorithm for the ANN model and having analysed the ED set, the optimal ANN model architecture and its parameter variation is determined. This is accomplished by selecting the optimum number of neurons in the hidden layer based on the minimum MSE value and the observed and predicted training and testing set values.

There is not a specific rule regarding the amount or percentage of data for training or testing and validation. The general guideline is that training data should be more than the testing and validation data [55, 68]. Hence, out of the total ED sets (i.e., 194,481) that was randomly divided by *trainlm*, 70% (i.e., 136,137) was used for

Table 1 ANNs experimental datasets generated for TRBN

	ANNs model inputs												ANNs model outputs		
Risk parameters	L			C			P			I			R		
Risk grades	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
1	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
2	1	0	0	1	0	0	1	0	0	0.8	0.2	0	0.95	0.05	0
3	1	0	0	1	0	0	1	0	0	0.8	0	0.2	0.95	0	0.05
4	1	0	0	1	0	0	1	0	0	0.6	0.4	0	0.9	0.1	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
194,478	0	0	1	0	0	1	0	0	1	0	0.6	0.4	0	0.15	0.85
194,479	0	0	1	0	0	1	0	0	1	0	0.4	0.6	0	0.1	0.9
194,480	0	0	1	0	0	1	0	0	1	0	0.2	0.8	0	0.05	0.95
194,481	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1

training, 15% (i.e., 291,72) for the testing, and 15% (i.e., 291,72) for the validation. In optimising the network, 15 neurons were used in the hidden layer as an initial trial, then the number of neurons was changed by increasing 10 neurons in each trial.

The preliminary trials indicated that the learning and prediction ability of 25 neurons in the hidden layer networks was better than that of 15 neurons. This was realised after several attempts to gradually increase the number of neurons and observe their effect on the predicted value. The training data error decreased, while that of validation data increased. Six local minimum MSE values were observed at neuron numbers of 15, 25, 35, 38, 40, and 45. However, the neural network architecture with 40 hidden neurons reached the minimum MSE when training, validating, and testing the ANN model. Thus, 40 neurons were chosen as the optimum number for the hidden layer. The network structure was 12–40-3 (12 neurons in the input layer, 40 in the hidden layer, and three in the output layer). The optimal ANN, together with a flowchart of the LMBP algorithm, is shown in Fig. 9. The training ended after 1000 iterations (*trainlm*, Epoch 1000) for the LMBP, when the differences between training and validation errors started to increase.

4.3 Results Validation

The network's MSE was very high, with 15 hidden neurons (MSE = 0.0524618), and it significantly decreased to a value of 0.0017914 with 25 hidden neurons. The number of neurons then increased from 25 to 35, and a gradual decrease of MSE was observed to a value of 0.00002151. Next, when the 35 hidden neurons increased to 45, MSE became 0.001394. 40 neurons were therefore tested, and the MSE reached its

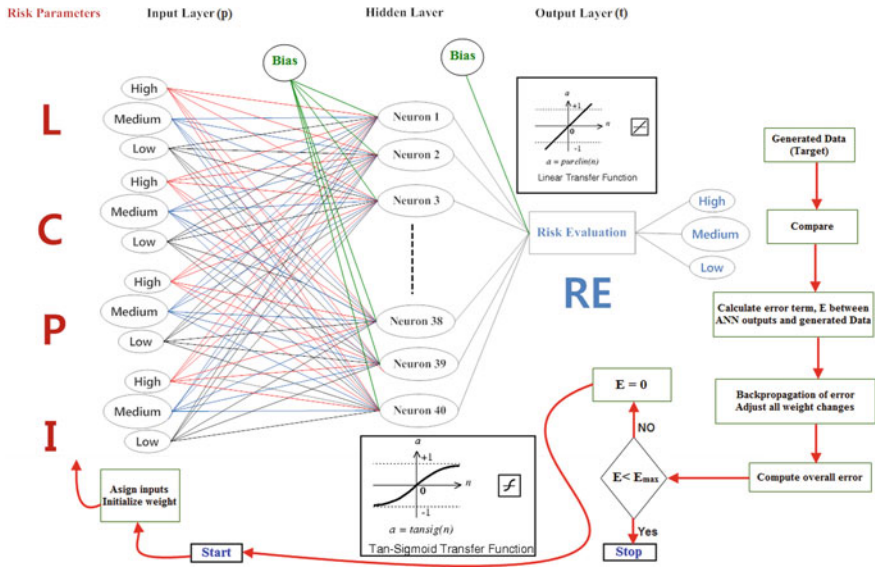


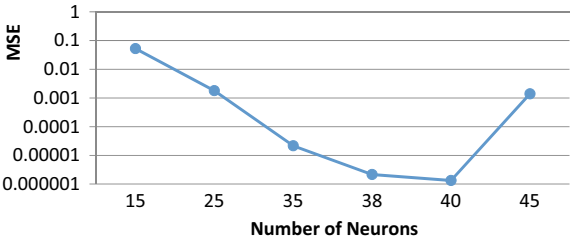
Fig. 9 Optimal FRBN-ANN structure with a flowchart of the LMBP

minimum value of 0.000001334. The neural network containing 40 hidden neurons (MSE 0.000001334) was chosen as the best case. When the number of neurons was less than 40, the MSE showed a slight increase from 0.000001334 to 0.000002164 at 38 neurons, as depicted in Fig. 10. This increment can be attributed to the characteristics of this study's MSE performance index and input vector, and it shows the dependence between MSE and number of hidden layer neurons for the LMBP.

The training, validation, and test's mean squared errors for the ANN using the LMBP algorithm are illustrated in Fig. 11. It shows that, with 40 neurons, the effect on training data error decreased while that on validation data increased.

The regression analysis of the network response between the ANN outputs and the corresponding targets was performed. The graphical output of the plotted network outputs versus the targets as open circles is illustrated in Fig. 12. Taking into account the data's non-linear dependence, linear regression shows a perfect agreement between the ANN outputs (predicted data) and the corresponding targets (i.e.,

Fig. 10 MSE for the FRBN-ANN tuning



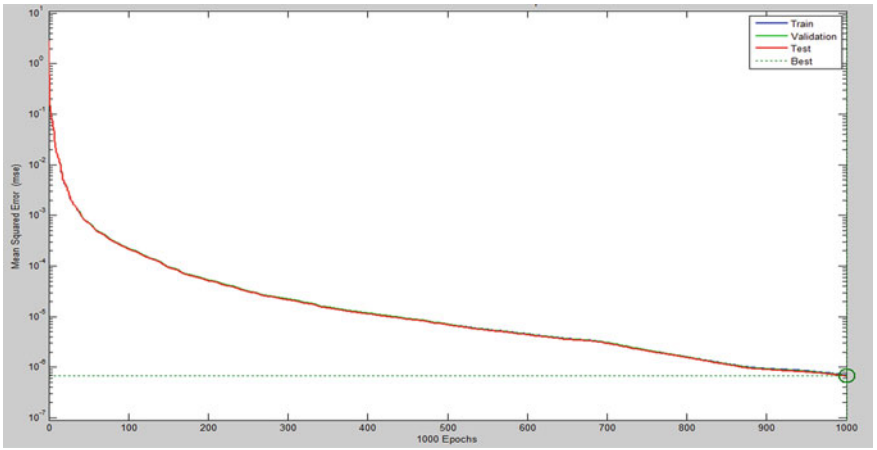


Fig. 11 Training, validation, and test's mean squared errors for the FRBN-ANN

ED). The solid red, blue, green, and black lines, representing the testing, training, validation, and the combination of all three respectively, indicate the perfect linear fit that $R^2 \approx 1$.

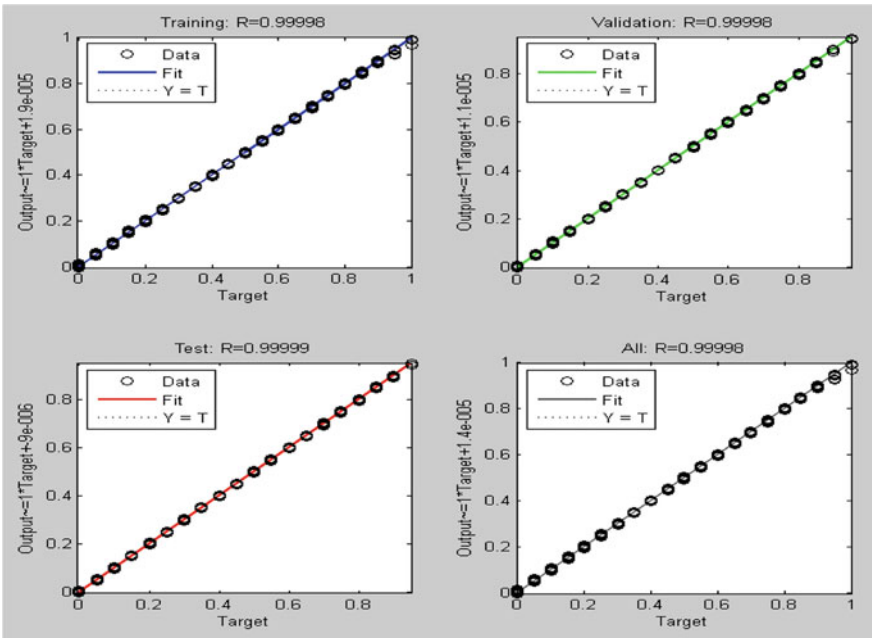


Fig. 12 The FRBN-ANN regressions

The previous statistical analysis shows that ANN model predictions are very close to the ED. In addition, to confirm the developed model's robustness and predictive capability, the optimal ANN model's performance was evaluated using another data set consisting of the actual data obtained from [20]. Consequently, a simulink model of the ANNs was constructed, and the results, along with the correlation coefficient between the actual and predicted datasets, shows a high accuracy and a perfect match.

5 The ANN Model Design to Simulate the Use of FER in Port Safety Evaluation

As aforementioned, FER model's inputs are taken directly from the FRBN model. Therefore, the ANN model for FER in this section is based on results directly taken from the ANN model for FRBN in Sect. 4. Similarly, the FER simulation using ANN is conducted through ED analysis, model optimisation and results validation.

5.1 ED Analysis of FER-ANN

This ED analysis is similar to the one in Sect. 4.1. However, the input parameters are the three grades (i.e., High, Medium and Low) of the three selected HEs (i.e., HE6, HE7, and HE 16). Therefore, the sum of all possible combinations of the DoB variations of the associated grades of the three HEs is calculated as 9261 (=21 × 21 × 21). Therefore, the EDs contained 9261 sets with nine inputs and one output in each set. This is obtained and partially shown in Table 2.

Table 2 The experimental datasets generated for the FER-ANN

No.	ANN model inputs									ANNs model outputs
	HE 1			HE 2			HE 3			
	High	Medium	Low	High	Medium	Low	High	Medium	Low	RI
1	1	0	0	1	0	0	1	0	0	1
2	1	0	0	1	0	0	0.8	0.2	0	0.9778
3	1	0	0	1	0	0	0.8	0	0.2	0.9556
4	1	0	0	1	0	0	0.6	0.4	0	0.9529
…	…	….	….	….	….	….	….	….	….	….
9258	0	0	1	0	0	1	0	0.6	0.4	0.075
9259	0	0	1	0	0	1	0	0.4	0.6	0.0471
9260	0	0	1	0	0	1	0	0.2	0.8	0.0222
9261	0	0	1	0	0	1	0	0	1	0

5.2 ANNs Model Optimisation

The same training algorithm used in Sect. 4.2 is used to optimise the ANN for FER (i.e., *trainlm*), including the same performance and assessment criteria with typical process and sequence of neurons number trails selection.

During several preliminary trial attempts to gradually increase the number of neurons and observe their effect on the predicted value, the training data error decreased and validation data error increased. It indicated that the learning and prediction ability of 25 neurons in the hidden layer networks was better than that of 15 neurons. Six local minimum MSE values were observed at neuron numbers of 15, 25, 35, 38, 40, and 45. However, the neural network architecture with 40 hidden neurons reached the minimum MSE. Consequently, 40 neurons were chosen as the optimum number for the hidden layer. Finally, the structure of the network was 9–40-1 (nine neurons in the input layer, 40 in the hidden layer, and one in the output layer). The LMBP training stopped at 51 iterations when the differences between training and validation error started to increase. In this process, out of the total datasets (i.e., 9,261) that was randomly divided by *trainlm*, 70% (i.e., 6483) was used for training, 15% (i.e., 1389) for testing, and 15% (i.e., 1389) for validation.

5.3 Results Validation

The network's MSE was very high for the 15 hidden neurons (i.e., 0.556) and decreased significantly from 25 to a value of 0.0000679. Then, as the number of neurons increased from 25 to 35, the value dropped to 0.0000617. Next, it increased again given that the hidden neurons increased from 35 to 45. Therefore, 40 neurons were tested, and the MSE reached at its minimum value of 0.00001344. Therefore, the neural network containing 40 hidden neurons was chosen as the best case, as seen in Fig. 13.

The FER associated ANN training, validation, and test's mean squared errors for using the LMBP algorithm are illustrated in Fig. 14. It clearly shows that, with 40 neurons, the effect on training data error's effect decreased, while that of validation data increased.

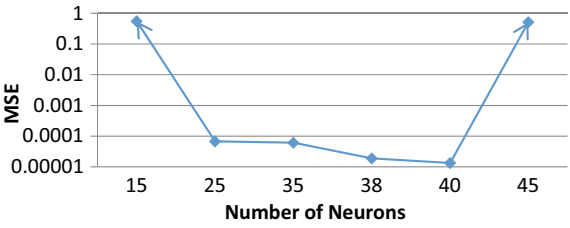


Fig. 13 MSE for FER-ANN optimisation

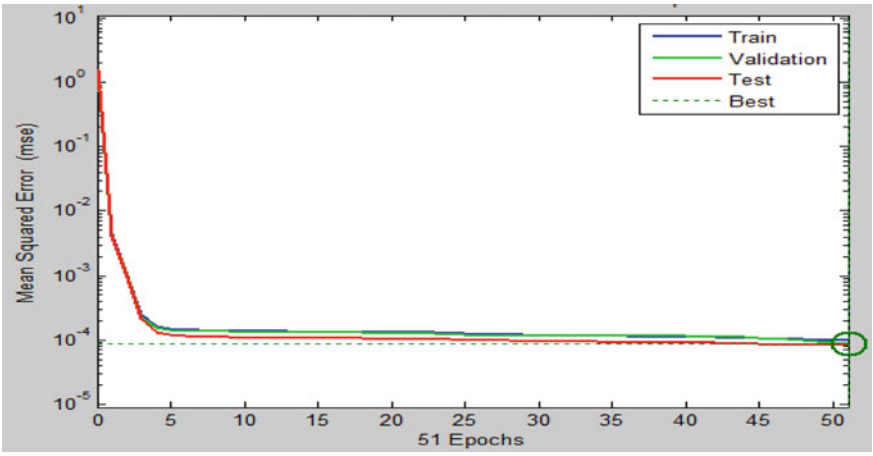


Fig. 14 Training, validation, and test mean's squared errors for the FER-ANN

The regression analysis of the network response between FER-ANN outputs and corresponding targets was performed. The graphical output of the plotted network outputs versus the targets as open circles is illustrated in Fig. 15. Taking into account the data's non-linear dependence, linear regression shows a perfect agreement between FER-ANN outputs (i.e., predicted data) and corresponding targets (i.e., ED). The solid red, blue, green, and black lines, representing the testing, training, validation, and the combination of all three respectively, indicate the perfect linear fit that $R^2 \approx 1$.

6 Implications

How the integrated FRBN-ANN and FER-ANN aids operators to enhance container port safety analysis under uncertainty.

The previous sections described the FRBN ANN and FER ANN models, which can be integrated to realise real time risk predication of a container port operation and provide a panoramic view on risk inference. The integration of the simulated BN and ER using ANNs was created within dynamic system simulation for MATLAB. Consequently, the simulink model of the integrated ANN network was constructed as illustrated in Fig. 16.

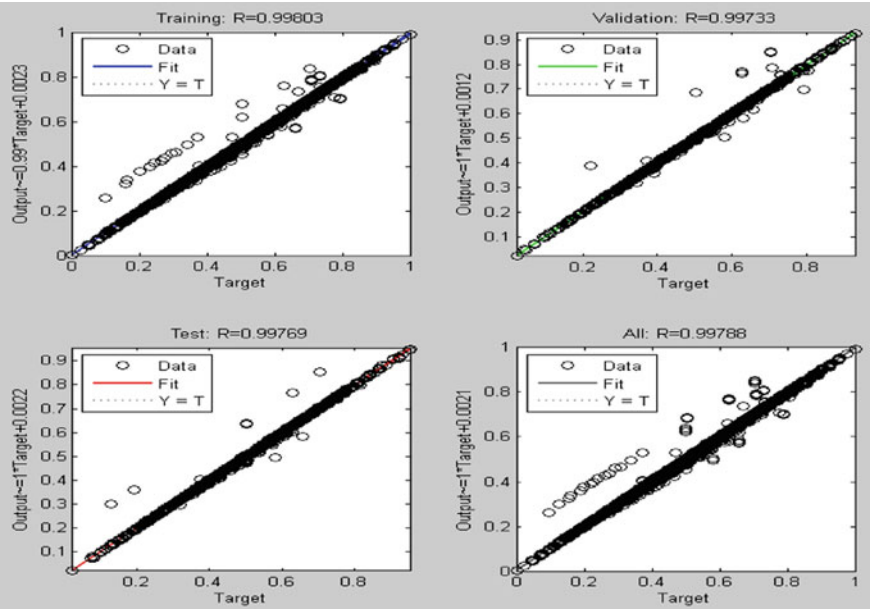


Fig. 15 FER ANN regressions

7 Conclusion

This chapter proposed a novel integrated ANN approach for evaluating and predicting the criticality of HEs in a container port system. It provides a panoramic view on container port risk analysis and prediction and useful insights for port safety management. Two three-layer BP neural networks were optimised to evaluate and predict the maritime container ports' operation safety. The configuration of the BP neural networks that generate the smallest MSE is conducted using the LMBP training algorithm of tansig transfer function at the hidden layer of 40 neurons and a purelin transfer function at the output layer. The optimal architecture for the FRBN-ANN and FER-ANN models were optimised as twelve and nine neurons in the input layers with three and one in the output layers, respectively. The two three-layer ANN-based models showed precise and effective predictions with satisfactory determination coefficients. The simulation of FRBN and FER using ANNs, with the presented results, showed that neural network modelling can effectively simulate and predict container terminal operation safety in one integrated framework. The developed software package using MATLAB can be used to facilitate the risk analysis, diagnosis and prediction by industrial stakeholders. They can obtain the updated safety performance of the investigated ports by simply providing the evaluations of the four risk parameters of any involved HE.

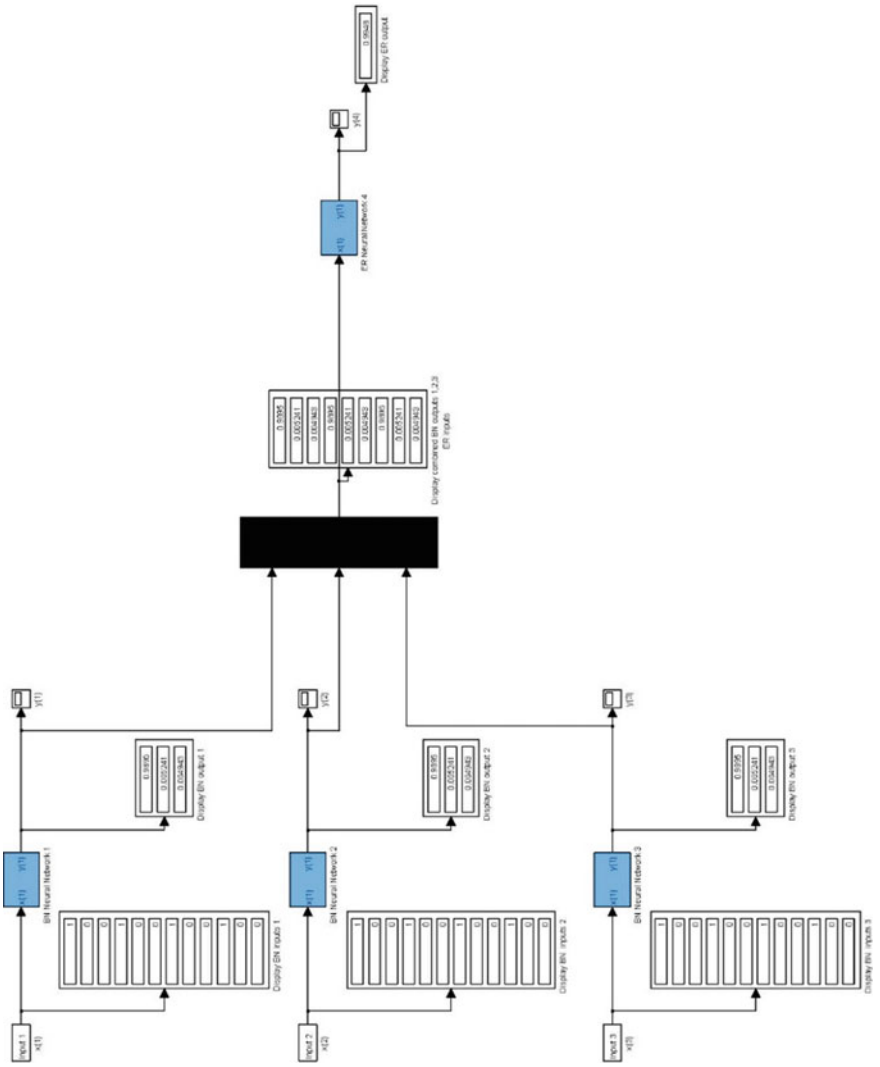


Fig. 16 Simulink model of the integrated ANN network

Furthermore, use of ANNs in FMEA can help overcome its incapability in tackling uncertainty in data and at the same time ease the evaluation process on the stakeholders ensuring the easiness of traditional risk priority number approach in FMEA. Although the aid from the associated software packages (i.e., Hugin for FRBN and IDS for FER), the separate FRBN and FER methods are still not user friendly for mathematically unsophisticated users. ANNs are proven to be effective to integrate different but correlated risk analysis methods collectively to deliver a reliable and robust risk prediction tool for container ports.

In the current study, the HEs are identified mainly from an operational perspective. Given the new ANN model's capability in handling high uncertainty in data, it will be feasible and beneficial to include other risk concerns influencing container port safety in future studies, including natural disasters, environmental, and political risks.

Acknowledgements This work was supported by the European Research Council under Grant [Number TRUST CoG 2019 864724].

References

1. Yang Z, Bonsall S, Wang J (2008) Fuzzy rule-based Bayesian reasoning approach for prioritization of failures in FMEA. IEEE Trans Reliab 57(3):517–528
2. Liu J, Yang JB, Wang J, Sii HS (2005) Engineering system safety analysis and synthesis using the fuzzy rule-based evidential reasoning approach. Qual Reliab Eng Int 21(4):387–411
3. Widrow B, Rumelhart DE, Lehr MA (1994) Neural networks: applications in industry, business and science. Commun ACM 37(3):93–106
4. Kumar PR, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. Eur J Oper Res 180(1):1–28
5. Sadatsafavi M, Moayyeri A, Soltani A, Larijani B, Nouraie M, Akhondzadeh S (2005) Artificial neural networks in prediction of bone density among post-menopausal women. J Endocrinol Invest 28:425–431
6. Wang X, Lederman D, Tan J, Wang XH, Zheng B (2011) Computerized prediction of risk for developing breast cancer based on bilateral mammographic breast tissue asymmetry. Med Eng Phys 33(8):934–942
7. Tsujita-Inoue K, Hirota M, Ashikaga T, Atobe T, Kouzuki H, Aiba S (2014) Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays. Toxicol In Vitro 28(4):626–639
8. Lacerda E, Carvalho AC, Braga AP, Ludermir TB (2005) Evolutionary radial basis functions for credit assessment. Appl Intell 22:167–181
9. Lin SL (2009) A new two-stage hybrid approach of credit risk in banking industry. Expert Syst Appl 36(4):8333–8341
10. De Andres J, Lorca P, de Cos Juez FJ, Sánchez-Lasheras F (2011) Bankruptcy forecasting: a hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). Expert Syst Appl 38(3):1866–1875
11. Oreski S, Oreski D, Oreski G (2012) Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Syst Appl 39(16):12605–12617
12. Odeyinka HA, Lowe J, Kaka AP (2013) Artificial neural network cost flow risk assessment model. Constr Manag Econ 31(5):423–439
13. Gomez H, Kavzoglu T (2005) Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin Venezuela. Eng Geol 78(1–2):11–27
14. Brack W, Bakker J, De Deckere E, Deerenberg C, Van Gils J, Hein M, Jurajda P, Kooijman B, Lamoree M, Lek S, López M, Marcomini A, Muñoz I, Rattei S, Segner H, Thomas K, Von P, Westrich B, De Zwart D, Schmitt M (2005). MODELKEY. Models for assessing and forecasting the impact of environmental key pollutants on freshwater and marine ecosystems and biodiversity. Environ Sci Pollut Res 12(5):252–256
15. Chen TY, Tsao CY (2008) The interval-valued fuzzy TOPSIS method and experimental analysis. Fuzzy Sets Syst 159(11):1410–1428
16. Moseley VJ, Dritsos SE, Kolaksis DL (2007) Pre-earthquake fuzzy logic and neural network based rapid visual screening of buildings. Struct Eng Mechan: An Int J 27(1):77–97

17. Wang YM, Elhag TM (2008) An adaptive neuro-fuzzy inference system for bridge risk assessment. Expert Syst Appl 34(4):3099–3106
18. Schuhmacher M, Nadal M, Domingo JL (2009) Environmental monitoring of PCDD/Fs and metals in the vicinity of a cement plant after using sewage sludge as a secondary fuel. Chemosphere 74(11):1502–1508
19. Ung ST, Williams V, Bonsall S, Wang J (2006) Test case based risk predictions using artificial neural network. J Safety Res 37(3):245–260
20. Alyami H, Lee PTW, Yang Z, Riahi R, Bonsall S, Wang J (2014) An advanced risk analysis approach for container port safety evaluation. Marit Policy Manag 41(7):634–650
21. Alyami H, Yang Z, Riahi R, Bonsall S, Wang J (2019) Advanced uncertainty modelling for container port risk analysis. Accid Anal Prev 123:411–421
22. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133
23. Basheer IA, Hajmeer M (2000) Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods 43(1):3–31
24. Lippmann R (1987) An introduction to computing with neural nets. IEEE ASSP Mag 4(2):4–22
25. Rosenblatt F (1961) Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Cornell Aeronautical Lab Inc Buffalo NY
26. Minsky M, Papert S (1987) Perceptrons: an introduction to computational geomety. 3rd edn. MIT Press Ed., Cambridge, pp 157–169
27. Marini F, Bucci R, Magrì AL, Magrì AD (2008) Artificial neural networks in chemometrics: history, examples and perspectives. Microchem J 88(2):178–185
28. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558
29. Liao SH, Wen CH (2007) Artificial neural networks classification and clustering of methodologies and applications–literature analysis from 1995 to 2005. Expert Syst Appl 32(1):1–11
30. Anderson J, Rosenfeld E (1988) In: Neurocomputing: foundations of research. MIT Press Ed., Cambridge, vol 6. pp 123–134
31. Pollack JB (1989) Connectionism: past, present, and future. Artif Intell Rev 3(1):3–20
32. McCord-Nelson M, Illingworth WT (1991) In: A practical guide to neural nets. Addison-Wesley Longman Publishing Co., Inc..
33. Eberhart RC (Ed.) (2014) In: Neural network PC tools: a practical guide. Academic Press
34. Priddy K, Keller P (2005) In: Artificial neural networks: an introduction. vol 68. SPIE Press
35. Yegnanarayana B (2009) In: Artificial neural networks. PHI Learning Pvt. Ltd.
36. Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS (2004) A neuro-fuzzy computing technique for modeling hydrological time series. J Hydrol 291(1–2):52–66
37. Kuo RJ, Chen CH, Hwang YC (2001) An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy Sets Syst 118(1):21–45
38. Caballero J, Fernández M (2008) Artificial neural networks from MATLAB® in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). Current Topics in Med Chem 8(18):1580–1605
39. Aggarwal P, Bhatt D, Devabhaktuni V, Bhattacharya P (2013) Dempster Shafer neural network algorithm for land vehicle navigation application. Inf Sci 253:26–33
40. Siegel PH, De Korvin A, Omer K (Eds) (1995) In: Applications of fuzzy sets and the theory of evidence to accounting, vol 2. Jai Press
41. Haykin S (1999) Neural networks: a comprehensive foundation. Prentice-Hall, New Jersey
42. Taha R (2012) The possibility of using artificial neural networks in auditing-theoretical analytical paper. European J Econ Finance Admin Sci 47
43. Dawes R (1991) In: Neural networks. vol 4(4). Addison-Wesley Publishing Company, Inc., New York,pp 531—533
44. Schalkoff R (1997) Artificial neural networks. McGraw- Hill, New York

45. Bar-Yam Y (2003) When systems engineering fails-toward complex systems engineering. In: SMC'03 conference proceedings. 2003 IEEE International conference on systems, man and cybernetics. Conference theme-system security and assurance (Cat. No. 03CH37483), October, vol 2. IEEE, pp 2021–2028
46. Simpson P (1990) Artificial neural systems: foundations, paradigms, applications, and implementations. Pergamon Press, New York
47. Maren A (1991) A logical topology of neural networks. In: Proceedings of the second workshop on neural networks, WNN-AIND 91. [Online]. Available http://www.aliannajmaren.com/Downloads/Logical-topology-neural-networks.pdf. Accessed 27 March 2014
48. Pham D (1994) Neural networks in engineering. In: Rzevski G et al. (eds) Applications of artificial intelligence in engineering IX, AIENG/ 94, Proceedings of the 9th international conference. Computational Mechanics Publications, Southampton, pp 3–36
49. Widrow B, Hoff ME (1960) In: Adaptive switching circuits. Stanford Univ Ca Stanford Electronics Labs
50. Blaško R (2000) Advances in intelligent and soft computing. Springer 5:233–237
51. MATLAB (2013) Neural network toolbox documentation. [Online]. Available: http://uk.mathworks.com/help/nnet/index.html. Accessed 21 March 2013
52. Statistica Neural Network (Statistica NN) (1998) Statistica NN Software Documentation, [Online]. Available: http://www.statsoft.com. Accessed 15 Nov 2015
53. Eren B, Ileri R, Dogan E, Caglar N, Koyuncu I (2012) Development of artificial neural network for prediction of salt recovery by nanofiltration from textile industry wastewaters. Desalin Water Treat 50(1–3):317–328
54. Bilgili M, Sahin B, Yasar A (2007) Application of artificial neural networks for the wind speed prediction of target station using reference stations data. Renew Energy 32(14):2350–2360
55. Kaveh NS, Ashrafizadeh SN, Mohammadi F (2008) Development of an artificial neural network model for prediction of cell voltage and current efficiency in a chlor-alkali membrane cell. Chem Eng Res Des 86(5):461–472
56. Magharei A, Vahabzadeh F, Sohrabi M, Rahimi KY, Maleki M (2012) Mixture of xylose and glucose affects xylitol production by Pichia guilliermondii: model prediction using artificial neural network
57. Mashhadi Meighani H, Dehghani A, Rekabdar F, Hemmati M, Goodarznia I (2013) Artificial intelligence vs. classical approaches: a new look at the prediction of flux decline in wastewater treatment. Desalination and Water Treatm 51(40–42):7476–7489
58. Yetilmezsoy K, Demirel S (2008) Artificial neural network (ANN) approach for modeling of Pb (II) adsorption from aqueous solution by Antep pistachio (Pistacia Vera L.) shells. J Hazardous Mater 153(3):1288–1300
59. Elmolla ES, Chaudhuri M, Eltoukhy MM (2010) The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. J Hazard Mater 179(1–3):127–134
60. Süt N, Çelik Y (2012) Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turkish J Med Sci 42(5):886–893
61. Kuo CFJ, Chiu HY, Syu SS, Vu QH (2014) The CO_2 laser parameter optimization design and practical verification for a touch panel conductive film. Opt Lasers Eng 52:250–260
62. Abushammala MF, Basri NEA, Elfithri R, Younes MK, Irwan D (2014) Modeling of methane oxidation in landfill cover soil using an artificial neural network. J Air Waste Manag Assoc 64(2):150–159
63. Najjar YM, Basheer IA, Hajmeer MN (1997) Computational neural networks for predictive microbiology: I methodology. Int J Food Microbiol 34(1):27–49
64. Jeffrey K, Chiu H, Syu S, Huy Q (2014) The CO_2 laser parameter optimization design and practical verification for a touch panel conductive film. Opt Lasers Eng 52:250–260
65. Hezave AZ, Lashkarbolooki M, Raeissi S (2012) Using artificial neural network to predict the ternary electrical conductivity of ionic liquid systems. Fluid Phase Equilib 314:128–133
66. Dogan E, Ates A, Yilmaz EC, Eren B (2008) Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand. Environ Prog 27(4):439–446

67. Turp S, Eren B, Ates A (2011) Prediction of adsorption efficiency for the removal of nickel(ii) ions by zeolite using artificial neural network (ANN) approach. Fresenius Environ Bull 20(12):3158–3165
68. Sudhakaran R, Vel Murugan V, Sivasakthivel PS, Balaji M (2013) Prediction and optimization of depth of penetration for stainless steel gas tungsten arc welded plates using artificial neural networks and simulated annealing algorithm. Neural Comput Appl 22:637–649
69. Lutz M (1996) In: Programming python. vol 8. O'Reilly
70. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 68:314–319

Maintenance Optimization and Warranty Policy

Usage of Failure Time and Repair Time for Optimization of Maintenance and Warranty Policy and Lemon Law Application

Minjae Park and Dong Ho Park

Abstract This chapter reviewed several optimization problems to determine its optimal relevant decision variables minimizing the expected warranty costs during certain intervals, such as life cycle, warranty period or post-warranty maintenance period. The decision variables of our interest include the length of warranty period, inter-PM interval and length of post-warranty maintenance period. All of the warranty models presented in this chapter are based on the renewable minimal repair-replacement (MRR) warranty under which both repair time and failure time are considered at the same time upon the system failure. Furthermore, the warranty conditions under the MRR warranty is somewhat similar to the ones regulated under the lemon law which aims to protect the buyers of the defective motor vehicles. The warranty model applicable to the lemon law is also presented in this chapter.

Keywords Failure time · Repair time · Lemon law · Maintenance · Warranty

1 Introduction

The subject of optimal warranty and maintenance policy for repairable system has been an important research issue in the field of reliability engineering for a long time and many researchers have studied and proposed a number of solutions for finding the best possible policies in the literature. In order to keep the repairable system in the operating state longer with a lower maintenance cost, it is a common practice for the manufacturer or the user to adopt an appropriate warranty and maintenance policy during the system's life cycle. This chapter discusses several optimization problems to find the best possible warranty and maintenance policies which minimize

M. Park
College of Business Administration, Hongik University, Seoul, South Korea
e-mail: mjpark@hongik.ac.kr

D. H. Park (✉)
Industry Academic Cooperation Foundation, Hallym University, Chuncheon, South Korea
e-mail: dhpark@hallym.ac.kr

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_12

the expected cost incurred for maintaining the system during its life cycle under a certain cost structure.

As for the warranty policy, there exist several types of warranties based on different conditions, such as number of dimensions, methods of compensation offered in a warranty contract, and renewable warranty or non-renewable warranty. One-dimensional warranty is based only on the failure time, but two-dimensional warranty policy considers both age and usage simultaneously to determine the warranty benefit and has been studied by many authors. As for the methods of compensations, free replacement/repair warranty(FRW), combination warranty(CMW) and pro-rata warranty(PRW) have been discussed in the literature. Under FRW and PRW, a failed system is repaired/replaced free of charge to the user and is charged prorated to the user in proportion to the usage of the system, respectively. CMW is a combination of both features of FRW and PRW, which often consists of two step warranties, the first one with a free warranty followed by a pro-rata warranty as a second step warranty. Under the renewable warranty, the warranty policy is renewed whenever the replacement occurs during the warranty period and the contract term is exactly the same as the original one. When the warranty is non-renewable, the system is warranted only during the original warranty period and once the warranty is expired, the system is warranted no further. Thus, the expected length of warranty period is a constant, not depending on the failures during the warranty period. Many different warranty models have been discussed in Blischke [1], Blischke and Murthy [2] and Park and Pham [3].

Maintenance actions are, in general, classified into two categories, preventive maintenance (PM) and corrective maintenance (CM). The PM is performed while the system is still in operating state and aims to slow down the system's degradation and the CM is carried out at the time of the system's failure and brings it back to an operational state. At each PM the system is inspected and upgraded to improve the performance by reducing its failure rate or age. Canfield [4] proposes a well-known failure rate reduction maintenance model, under which each PM lowers the failure rate to that existing during certain time units prior to the current PM time. Later, Kim et al. [5] suggest another type of failure reduction model adopted at each PM, where the failure rate is adjusted somewhat lower and the level of reduction is determined by an improvement level. More discussions on PM is given in Sect. 3.

In Sect. 2 we study a renewable minimal repair-replacement warranty strategy, which relies on both failure time and length of repair time simultaneously for the failed system. Under such a two-dimensional warranty policy, a repair time threshold is pre-determined and if the repair work for the failed system can't be finished within the repair time threshold, the system is replaced by a new one. By setting the threshold necessary to repair the system failures, the warranty can protect the user more effectively from a long wait for the completion of minimal repair on each system failure, in spite of the cost increase from the manufacturer's perspective. Such a two-dimensional warranty model originally has been proposed in Park et al. [6]. Other type of two-dimensional warranty utilizing two factors, age and usage, has been considered by many researchers in studying the warranty policy. In this warranty model, the age of the system can be observed easily, but the usage could

be quite difficult to obtain in many situations. However, the failure time and length of repair time can be readily obtainable on each failure of the system and thus the two-dimensional warranty model based on both failure time and repair time can be more beneficial for the user to adopt due to the accessibility of such data. The two-dimensional warranty model we analyze in this chapter works as follows. When the system stops operation or malfunctions during the warranty period, the manufacturer would provide the minimal repair service first. If the minimal repair can be finished within the pre-determined repair time threshold, the repaired system is returned back to the operation and the warranty remains unchanged as it is. However, in case the minimal repair can't be finished within the repair time threshold, then the failed system is replaced with a new one and the warranty is renewed. Thus, if a certain number of replacements occur during the warranty period, the warranty is extended proportionally to the number of replacements. If no replacement occurs during the warranty period, the current warranty period stands with no further extension.

Another important aspect of warranty model we consider in this chapter is regarding the lemon law, which is being enforced in many countries. This chapter also discusses an optimal maintenance policy applicable under the lemon law, which is originally for the motor vehicles being the major target. The lemon law aims to protect consumers from purchasing defective systems, which meet certain conditions regarding the number of failures or the accumulated repair time for a new system during a pre-specified period. Such a period is in general within the warranty period.

A "lemon" is defined as a defective system and while the conditions for declaring a system as a lemon are generally similar in most of enforcing countries, they differ slightly in several details depending on the state or the country. For example, in California, a system is determined to be a lemon if the accumulated repair time takes more than 30 days or if the number of failures exceeds a pre-determined failure number threshold. The lemon law obligates the manufacturer or the seller to refund all or part of the purchasing price to the user or replace the lemon with a new one. Some references dealing with lemon laws specifically include Goldberg and Paz [7], Snyder and Daskin [8], Iskandar and Husniah [9], Wang et al. [10] and Park et al. [11, 12]. The optimal maintenance and warranty policies applicable to the lemon law are presented in Sect. 5.

This chapter presents several types of optimal warranty and maintenance models by determining the relevant decision variables, which minimize the expected warranty and maintenance costs incurred during certain intervals such as the life cycle of the system or the warranty period. The decision variables of our interest to be determined in this chapter include length of warranty period, length of post-warranty maintenance period, length of inter-PM interval during the warranty period, etc. In Sect. 2 some notations and assumptions are given and the renewable minimal repair and replacement warranty model is described in details. Sections 3 and 4 present several optimal maintenance policies incorporating the renewable repair and replacement warranty model from the perspectives of both manufacturer and user. A few possible research issues of interest are discussed in concluding remarks in Sect. 6.

2 Two-Dimensional Minimal Repair-Replacement Warranty Model

The search for an optimal warranty policy is a critical issue for the manufacturer to make a decision to reduce the possibility of the system failures during the warranty period which in turn lowers the warranty cost. On the other hand, the user is charged a certain portion of warranty cost even during the warranty period in most situations. Furthermore, once the warranty expires, the maintenance of the system becomes a sole responsibility of the user. In this regard, the user is concerned to find the best possible warranty and maintenance strategy during the life cycle of the system, especially during the post-warranty maintenance period. Due to such necessity to develop an optimal warranty and maintenance policy, a large number of works have been done and proposed in the literature from the perspectives of both manufacturer and user. Nakagawa [13] is an excellent article to refer for such research works.

A warranty is an obligation that requires the manufacturer to provide compensation for the user according to the warranty terms in case the warranted system fails to conform to their intended functions. From the user's perspective, a warranty which is offered by the seller or the manufacturer at the time of purchasing the system provides a certain degree of protection for the user in case the system failures occur due to the low reliability of the system or the system's poor performance. On the other hand, the warranty also helps the manufacturer as well since the warranty gives a kind of limitation for the manufacturer's responsibility explicitly in terms of length of warranty and type of compensation upon the system failures. Recently, it appears that the warranty needs to be more comprehensive and becomes costly due to complexity of the system and intense competitions among the competing manufacturers and thus the warranty becomes longer and covers wider range of system failures. Consequently, many researchers have worked on optimal maintenance polices incorporating various types of warranty models in the last several decades and there exist a number of maintenance and warranty models proposed in the literature.

In this section, we discuss a warranty model which is affected by two factors of failure time and repair time simultaneously when the system failures occur during the warranty period.

2.1 Assumptions and Notations

The following assumptions and notations are commonly used throughout this chapter.

2.1.1 Assumptions

- The system is assumed to be repairable and deteriorating as it ages.
- All the warranty claims are valid and accepted.

- All repairs are assumed to be minimal and the failure process of the system follows the NHPP.
- Minimal repair cost is free of charge to the user in the warranty period.
- Necessary times for replacement and minimal repair of the failed system is excluded from the warranty period.
- Necessary time for the PM action is assumed to be negligible.
- Once the warranty expires, the system is maintained solely by the user and all post-warranty maintenance costs are charged to the user.

2.1.2 Nomenclature and notations

Nomenclature

- NHPP: Non-Homogeneous Poisson Process
- $ECR_{(\cdot)}$: Expected Cost Rate
- MRR: Minimal Repair-Replacement
- cdf, pdf: cumulative distribution function, and probability density function, respectively
- r.v.: random variable
- i.i.d.: independent and identically distributed
- PM: Preventive Maintenance
- T, Y: random variables representing failure time and repair time of the system, respectively.

Notation

- $\lambda(\cdot)$: intensity function of the NHPP for the system failure
- $f(\cdot)$, $F(\cdot)$, $\overline{F}(\cdot)$: pdf, cdf and reliability function of failure time T, respectively
- $g(\cdot)$, $G(\cdot)$, $\overline{G}(\cdot)$: pdf, cdf and reliability function of Y, respectively
- $h(t)$: failure rate function of T
- $f_{pm}(\cdot)$, $F_{pm}(\cdot)$, $h_{pm}(\cdot)$: pdf, cdf and failure rate function of T with PM adjustment
- δ_m: fixed length of post-warranty maintenance period
- δ: PM period between two successive periodic PM actions
- α: PM restoration level
- Ψ: warranty region which is censored by both repair time threshold and warranty period
- C_r, C_m, $C_{pm,}C_f$: random variables representing total replacement cost, total minimal repair cost, total PM cost, and total failure cost, respectively
- c_r, c_m, $c_{pm,}c_f$: unit replacement cost, unit minimal repair cost, unit PM cost, and unit failure cost, respectively
- C_c, C_n, C_{re}: random variable representing warranty cost due to critical and non-critical component failures and refund cost, respectively, under the lemon law
- c_c, c_n : unit minimal repair cost for critical and non-critical component failures, respectively.
- N_R: number of replacements during the warranty period under renewable MRR

- N_T: total number of system failures in the warranty period under renewable MRR
- N_Ψ: number of minimal repairs without PM in the warranty period under renewable MRR
- N_M: number of minimal repairs conducted during the warranty period when PM actions are taken.
- N_{np}: total number of system failures during the post-warranty maintenance period
- r_0: repair time threshold
- M: number of PMs conducted in the post-warranty period
- δ_p: length of inter-PM interval during the post-warranty period
- ω, ω_0: length of original warranty period and extended warranty period, respectively
- $EC_{(\cdot)}(\omega, \delta)$: expected total warranty cost when the warranty period equals ω and the optimal PM interval equals δ
- $H_{\mathrm{m}}\left(\omega, \delta_p\right)$: expected number of minimal repairs occurred during the post warranty period
- I_j: inter-replacement time interval elapsed between the $(j-1)st$ and the j th replacement during the warranty period, $\mathrm{j} = 1, 2, \cdots \mathrm{N_R}$
- $m(\omega, r_0)$: mean length of I_j, $\mathrm{j} = 1, 2, \cdots$.
- $\lfloor x \rfloor$: integer part of number x

2.2 *Renewable Minimal Repair and Replacement*

The warranty policy we study in this section is a renewable two-dimensional warranty policy with a pre-specified repair time threshold. Many authors, including Iskandar et al. [14], Jung and Bai [15], Chen and Popova [16], and Ye et al. [17] among many others, have considered two-dimensional warranty policies based on the system's age and usage when the product's failure occurs during the warranty period. Blischke and Murthy [2] also develop a two-dimensional warranty in terms of usage and age. Unlike these approaches, a warranty model using the repair time and failure time as two factors was first proposed by Park et al. [6] and was utilized to develop the optimal post-warranty maintenance policy. Such a two-dimensional warranty is referred to as a renewable minimal repair-replacement (MRR) warranty throughout this chapter. Park et al. [6] assert that two factors of repair time and failure time are easier to access than the traditional two factors of system age and usage in practice.

2.2.1 Description of Renewable MRR Model

Consider a renewable MRR model with the original length of warranty period, denoted by ω, and the failure time threshold, denoted by r_0. Under such a warranty model, the manufacturer is responsible for providing the replacement/minimal repair services upon the system failures in the warranty period. When the repair work exceeds the repair time threshold, a replacement is provided and the warranty policy

is renewed for the replaced system with exactly the same warranty terms as the original ones.

The graphical representation for the renewable MRR model is shown in Fig. 1. For j $= 1, 2, \cdots,$ we let $I_j (< \omega)$ denote the inter-replacement time interval elapsed between the $(j-1)st$ and the j th replacement of the system during the warranty period. In Fig. 1, if $I_1 < \omega$, then under the renewable MRR model, the replacement service for the failed system will be carried out by the manufacturer and the replacement cost will be either free or pro-rated to the user depending on the type of MRR warranty policy. Starting from I_1, the system which has been replaced will have the same renewable warranty with the length of period of ω again as shown in Fig. 1. Since the warranty is extended by ω each time the replacement occurs, the length of actual valid warranty period becomes a random variable (*r.v.*). Let W_0 denote the length of warranty cycle which is defined as the time interval starting from the purchasing time and ending at the expiration of warranty. It is clear that for a non-renewable MRR warranty, a warranty cycle coincides with the warranty period of ω. However, for a renewable MRR model, W_0 is a *r.v.* depending on the total number of system replacements, the inter-replacement times between two successive replacements in the warranty period and the original length of warranty period. Let $N_T = N_\Psi + N_R$ be the total number of system failures in the warranty period, where N_R and N_Ψ are the number of replacements and the number of minimal repairs, respectively. Then, W_0 can be expressed as a function of ω as follows.

$$W_0 = L(\omega) = I_1 + I_2 + \cdots + I_{N_R} + \omega \tag{1}$$

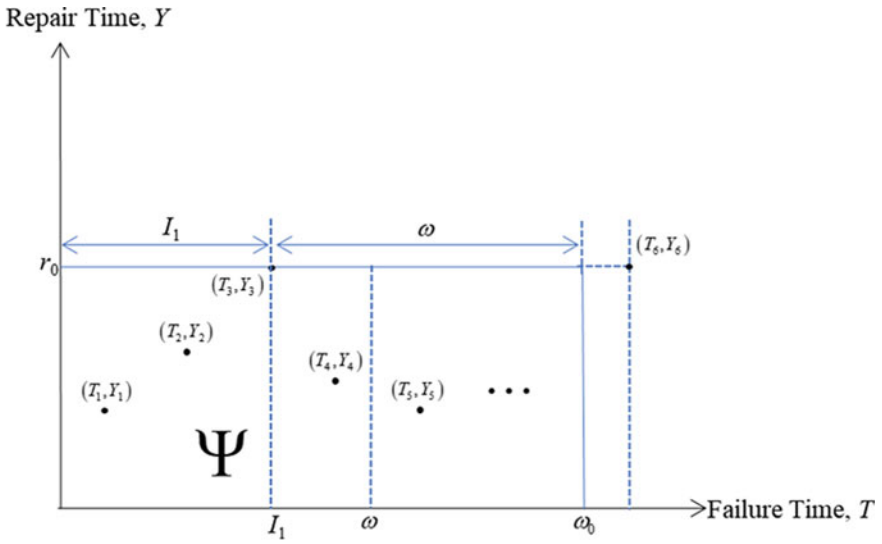


Fig. 1 Diagram for renewable MRR model

Figure 1 describes the case for which the first replacement occurs at time I_1 and the second replacement is not provided until the warranty expires. It is clear that $\omega_0 = I_1 + \omega$ where ω_0 is the realization of W_0. In Fig. 1, T_j and Y_j, $j = 1,2,\cdots$ denote the time at which the jth failure occurs and the length of its corresponding repair time, respectively. Let Ψ denote the warranty region which is censored by the warranty period and the repair time threshold.

This section mainly considers the probabilistic aspects of the two-dimensional warranty model, which is affected by both the time of system failure and the length of its corresponding repair time. For such purpose, the notion of repair time threshold is newly introduced as a criterion to determine whether the failed system would be replaced or minimally repaired during the warranty period. In this context, we formulate the cost model incorporating the two-dimensional warranty policy under study. Then we evaluate the manufacturer's expected warranty cost incurred during the warranty period and decide the optimal warranty length so as to minimize the expected warranty cost.

2.2.2 Length of Warranty Period

To derive a formula to evaluate the expected cost rate (ECR) incurred in the warranty period under the renewable MRR warranty model, we need to find the expected length of warranty period by deriving the probability distribution of $L(\omega)$, defined in Eq. (1). Assume that the system is replaced N_R times in the warranty period. Though the system has additional failures during the warranty period which need only minimal repairs, these failures have no effects on the length of warranty period. Note that the warranty is renewed to extend the warranty period only when the replacement occurs. By adding the lengths of inter-replacement intervals, the extended warranty length becomes equal to $L(\omega) = \sum_{j=1}^{N_R} I_j + \omega$, and thus the expected length of warranty period, conditioned on $N_R = n$, can be obtained as

$$\begin{aligned} \mathrm{E}(L(\omega)|N_R = n) &= \sum_{j=1}^{N_R} E\left(I_j + \omega | N_R = n\right) \\ &= \sum_{j=1}^{N_R} E\left(T_j | T_j \le \omega, Y_j \ge r_0, N_R = n\right) + \omega \end{aligned} \tag{2}$$

Replacement by a new system is given by the manufacturer only when the system failure occurs in the warranty period and its repair time exceeds the pre-specified repair time threshold. Thus, the total number of system replacements in the warranty period has the following geometric distribution.

$$\mathrm{P}(N_R = n) = \left\{\overline{F}(\omega) + F(\omega) \cdot G(r_0)\right\} \cdot \left\{F(\omega) \cdot \overline{G}(r_0)\right\}^n, \quad n = 0, 1, 2, \cdots. \tag{3}$$

Taking the expectation on $\mathrm{E}(L(\omega)|N_R = n)$, given in Eq. (2), with respect to N_R, we obtain the expected warranty length as

$$\begin{aligned}\mathrm{E}(L(\omega)) &= \sum_{j=0}^{\infty} \left\{\overline{F}(\omega) + F(\omega)\cdot G(r_0)\right\} \\ &\quad \cdot \left\{F(\omega)\cdot \overline{G}(r_0)\right\}^j \cdot \left\{ j\cdot \frac{\int_0^{\omega} t\cdot f(t)dt}{F(\omega)\cdot \overline{G}(r_0)} + \omega \right\} \\ &= \int_0^{\omega} t\cdot f(t)dt \Big/ \left\{1 - F(\omega)\cdot \overline{G}(r_0)\right\} + \omega \end{aligned} \tag{4}$$

by applying Eq. (3).

The expected warranty length under the renewable MRR, given in Eq. (4), can be used to formulate the expected cost rate (ECR), which is to be used as an objective function for optimization problem in the subsequent sections.

3 Optimal Warranty Policy from the Manufacturer's Perspective

It is a common practice in the market that at the sale of the system, the manufacturer or the seller offers a certain type of warranty to the user regarding the maintenance of the system under the manufacturer's responsibility during a certain length of time. From the manufacturer's perspective, a profitable and attractive warranty policy is an important issue not only for enhancing the system's sale, but also for reducing the warranty cost incurred during the warranty period. In this regard, many researchers have studied various types of warranty policies along with the maintenance strategy during the entire life cycle of the system for a long time. In general, the life span of the system begins with the installation of a new system and ends when it is replaced by another system. One of the main purposes of maintenance policy is to minimize the maintenance cost incurred either during the life cycle of the system or during the warranty period from the manufacturer's point of view, while keeping the product at maximum availability. The effectiveness of a maintenance policy may count on the nature of the warranty policy. In this respect, many studies have been carried out on the subject of optimal maintenance policy, such as those by Park and Pham [3] and Shafiee and Chukova [18], who discuss various warranty policies and their related optimization aspects.

The warranty policy can be classified into several types of warranties based on different conditions, such as free or pro-rata warranty, renewable or non-renewable warranty, and so on. Another way of differentiating the warranty is either one-dimensional or two-dimensional warranty, which is determined by the number

of factors affecting the warranty. The warranty is usually referred to as a two-dimensional when two factors, such as failure time and usage, affect the warranty, while one-dimensional warranty is dependent only on the time of system failure.

In this section, we present two optimal warranty policies for the manufacturer to offer by considering the renewable MRR warranty, which is discussed in Sect. 2. For that purpose, the PM action is briefly discussed in this section. The PM is a maintenance action taken while the system is still in operation and is known to be very effective scheme to keep the system in an operating state longer and is most widely used to maintain the system. Under the renewable MRR warranty, the lengths of the warranty period and the inter-PM interval in case the periodic PM actions are warranted by the manufacturer in the warranty period are two major decision variables for the manufacturer to make a decision for its optimization. This section presents the optimal warranty policies by determining these two variables so that the expected cost rate incurred is minimized.

3.1 Preventive Maintenance During Warranty Period

A periodic PM is a well-known preventive action to slow down the deterioration of the repairable system, while the system is still in operation. The periodic PM action inspects the performance of the system and improves the system's reliability periodically by lowering its failure rate or age to some extent. The manufacturer of motor vehicles usually provides the periodic oil and filter changes for the vehicle, new or used, for a certain length of period, which can be considered as a good practical example for such a periodic PM action. Although the frequent PMs increase the maintenance cost during the life span of the system, such actions may help to lower the chance of system failures by improving the operating conditions of the system. In this regard, the optimal trade-off between the maintenance cost and the number of PM actions becomes an important subject to solve for both the manufacturer and the user. As a result, many researchers have been studying to propose a number of optimal PM policies in the literature. Canfield [4] proposes a periodic maintenance model, under which the failure rate of the system is reduced at each PM action. Later, Shafiee et al. [19] also discuss optimal PM warranty strategies with respect to the level of PM, the PM interval, and the number of PM actions. Many other references for optimal maintenance models considering PM actions can be found therein. Two-dimensional warranty utilizing age and usage has also been discussed in Huang et al. [20], Wang et al. [21, 22], Huang et al. [23], Su and Wang [24] and others, regarding the PM policies for the repairable systems. Besides, there exist several optimal maintenance policies incorporating the PM actions for the second-hand system as well. To mention a few of them, Lim et al. [25], Khatab et al. [26], Yeh et al. [27], Su and Wang [28] and many others have considered the periodic PM actions to study the optimal maintenance problem for the second-hand product.

Canfield [4] has proposed a failure reduction model, which has been widely used by researchers to improve the system performance at each PM. In Canfield's model,

each PM lowers the failure rate to that existing during certain time units prior to the current PM time. Such time units are referred to as the level of restoration that is less than or equal to the PM interval. The failure rate function of the system with the PM action under Canfield's [4] model can be expressed as

$$h_{pm}(t) = \begin{cases} h(t) & for\ 0 \le t \le \delta \\ \sum_{i=1}^{k} \{h((i-1)(\delta-\alpha)+\delta) - h(i(\delta-\alpha))\} & for\ k\delta < t \le (k+1)\delta, k = 1, 2, \cdots, \\ +h(t-k\alpha) & \end{cases} \tag{5}$$

where $h(t)$ denotes the failure rate of the system before the first PM is conducted and δ is the length of inter-PM interval between two successive periodic PMs. Here, α is a measure to indicate the level of restoration with $0 \le \alpha \le \delta$, which measures the PM effect. Note that $\alpha = 0$ implies no PM effect and $\alpha = \delta$ implies the perfect PM effect.

Recently, Kim et al. [5] propose other type of PM model, which reduces the current failure rate to a certain extent determined by an improvement level, to study an optimal maintenance policy for a second-hand system. Under Kim et al.'s [5] model, the failure rate of the system can be adjusted to

$$\mathrm{h}_{\alpha}(\mathrm{t}) = \begin{cases} h(\mathrm{t}), & x \le \mathrm{t} < x + \xi \\ \mathrm{k}\alpha\{h(\mathrm{x}+\xi) - \mathrm{h}(\mathrm{x})\} + \mathrm{h}(\mathrm{t} - \mathrm{k}\xi), & x + k\xi \le \mathrm{t} < x + (k+1)\xi \end{cases} \tag{6}$$

after k periodic PM actions are taken for $k = 1, 2, \cdots$. Here α denotes the improvement level and as the value of α is taken smaller, the reduction of failure rate becomes larger. Thus, $\alpha = 0$ implies that the failure rate is reduced to that of a new system and $\alpha = 1$ implies that no improvement is made at all.

3.2 *Optimal Length of Warranty Period Under Renewable MRR Model*

An effective warranty policy becomes an important issue at the sale of the system for the manufacturer in terms of its sales volume and warranty cost. A certain kind of warranty is offered to the user whenever the transaction of the system is made in most situations and the contract terms of warranty affect not only the sale of the system, but also the warranty cost incurred in the warranty period from the manufacturer's perspective. In this respect, many authors have worked on the optimization of maintenance strategy incorporating the warranty policy for a repairable system in the last several decades.

The main objective of this subsection is to determine an optimal length of warranty period minimizing the warranty cost under the renewable MRR warranty model, discussed in Sect. 2. Most recently, Park et al. [29] propose an optimal warranty policy under the renewable MRR warranty model by finding an optimal length of warranty period minimizing the expected cost rate per unit time in the warranty period. Park et al. [29] discuss such an optimization problem for both free warranty and pro-rata warranty for replacement cost incurred in the warranty period.

Let C_r, C_m and C_f denote the *r.v.'s* representing total costs for replacement, minimal repair, and failure, respectively, during the warranty period from the manufacturer's perspective. Then the expected total warranty cost can be obtained as $\mathrm{E}(C_r) + \mathrm{E}(C_m) + \mathrm{E}\big(C_f\big)$. Firstly, we evaluate the expected replacement cost under the assumption that the system is replaced by a new one N_R times during the warranty period. Under the renewable pro-rata MRR, the replacement cost, denoted by C_r, is evaluated as a function of I_j as follows.

$$\mathrm{C_r} = c_r \cdot \sum_{j=1}^{N_R} \left(1 - \frac{I_j}{\omega}\right).$$

In case of a renewable free MRR model, $\mathrm{C_r} = c_r \cdot N_R$. Throughout this section, we consider only the renewable MRR model with pro-rata replacement cost. Let T_j denote the jth failure time of the system. Then, the expected replacement cost, conditioned on $N_R = n$, can be evaluated as

$$\begin{aligned}
\mathrm{E}(C_r|N_R = n) &= c_r \left\{ E\left(\sum_{j=1}^{N_R} \left(1 - \frac{I_j}{\omega}\right) | I_j \le \omega, N_R = n \right) \right\} \\
&= c_r \left\{ n - \frac{1}{\omega} \cdot \sum\nolimits_{j=1}^{n} E\big(T_j | T_j \le \omega, Y_j \ge r_0\big) \right\} \\
&= c_r \left\{ n - \frac{n}{\omega} \cdot \frac{\int_0^{\omega} t \cdot f(t)dt}{F(\omega) \cdot \overline{G}(r_0)} \right\} \qquad (7)
\end{aligned}$$

Note that N_R has the geometric distribution, as given in Eq. (3). Thus by taking the expectation on $\mathrm{E}(C_r|N_R = n)$ of Eq. (7) with respect to N_R, the expected total replacement cost is obtained as

$$\begin{aligned}
&\mathrm{E}(C_r) = E(\mathrm{E}(C_r|N_R = n)) \\
&= \sum_{j=0}^{\infty} \left\{ \bar{F}(\omega) + F(\omega) \cdot G(r_0) \right\} \cdot \left\{ F(\omega) \cdot \bar{G}(r_0) \right\}^j \\
&\cdot \left\{ j \cdot c_r \cdot \left(1 - \frac{\int_0^{\omega} t \cdot f(t)dt}{\omega \cdot F(\omega) \cdot \bar{G}(r_0)} \right) \right\}
\end{aligned}$$

$$= c_r \cdot \left(\frac{\omega \cdot F(\omega) \cdot \bar{G}(r_0) - \int_0^{\omega} t \cdot f(t)dt}{\omega\big(1 - F(\omega) \cdot \bar{G}(r_0)\big)} \right) \tag{8}$$

Next, we consider the expected minimal repair cost and failure cost, which are free to the user. The system is minimally repaired for its failure which occurs during the warranty period and the repair work can be completed within the repair time threshold r_0. Let N_Ψ denote the number of minimal repairs in the warranty period. Then the minimal repair cost can be expressed as $C_m = c_m \cdot N_\Psi$. Since N_Ψ can be counted as

$$N_\Psi = \int_0^{I_1} h(t)dt + \int_0^{I_2} h(t)dt + \cdots + \int_0^{I_{N_R}} h(t)dt + \int_0^{\omega} h(t)dt,$$

for given $N_R = n$, the conditional expectation of C_m can be obtained as

$$\mathrm{E}(C_m | N_R = n) = c_m \cdot \left(n \cdot \int_0^{A} h(t)dt + \int_0^{\omega} h(t)dt \right), \tag{9}$$

where $\mathrm{A} = \frac{\int_0^{\omega} t \cdot f(t)dt}{F(\omega) \cdot \overline{G}(r_0)}$. By taking the expectation again on $\mathrm{E}(C_m | N_R = n)$ of Eq. (9) with respect to N_R, the expectation for total minimal repair cost can be evaluated by the following equation.

$$\mathrm{E}(C_m) = c_m \cdot \left(\frac{F(\omega) \cdot \overline{G}(r_0)}{1 - F(\omega) \cdot \overline{G}(r_0)} \cdot \int_0^{A} h(t)dt + \int_0^{\omega} h(t)dt \right) \tag{10}$$

When the system failure occurs, we assume that a certain amount of failure cost may occur and is paid by the manufacturer. Since the total number of failures in the warranty period, denoted by N_T, equals $N_T = N_R + N_\Psi$, the total failure cost can be obtained as $C_f = c_f \cdot (N_R + N_\Psi)$. By adding the replacement cost, minimal repair cost and failure cost, we obtain the total warranty cost from the manufacturer's perspective as

$$\mathrm{C}(\omega) = C_r + C_m + C_f = c_r \cdot \sum_{j=1}^{N_R} \left(1 - \frac{I_j}{\omega} \right) + c_m \cdot N_\Psi + c_f \cdot (N_R + N_\Psi). \tag{11}$$

By taking the expectation on $\mathrm{C}(\omega)$ given in Eq. (11), it can be shown that the expected total warranty cost can be expressed as

$$\mathrm{E}(\mathrm{C}(\omega)) = c_r \cdot \left(\frac{\omega \cdot F(\omega) \cdot \overline{G}(r_0) - \int_0^{\omega} t \cdot f(t)dt}{\omega\big(1 - F(\omega) \cdot \overline{G}(r_0)\big)} \right) + c_f \cdot \left(\frac{F(\omega) \cdot \overline{G}(r_0)}{1 - F(\omega) \cdot \overline{G}(r_0)} \right)$$

$$+\left(c_m + c_f\right) \cdot \left(\frac{F(\omega) \cdot \overline{G}(r_0)}{1 - F(\omega) \cdot \overline{G}(r_0)} \cdot \int_0^A h(t)dt + \int_0^\omega h(t)dt \right). \tag{12}$$

For more detailed discussions on derivation of $\mathrm{E}(\mathrm{C}(\omega))$, you may refer to Park et al. [29]. To determine the optimal length of warranty period under the renewable MRR warranty model, we utilize the expected cost rate(ECR) per unit time in the warranty period. Denoting the ECR under our proposed warranty model by $ECR_{(1)}(\omega)$, we have

$$ECR_{(1)}(\omega) = \frac{\mathrm{E}(\mathrm{C}(\omega))}{\mathrm{E}(L(\omega))}, \tag{13}$$

where $\mathrm{E}(L(\omega))$ and $\mathrm{E}(\mathrm{C}(\omega))$ are given in Eqs. (4 and 12), respectively.

By utilizing $ECR_{(1)}(\omega)$, given in Eq. (13), as an objective function to find an optimal solution for ω, denoted by ω^*, the optimization problem can be formulated as

$$\text{Find } \omega^*, \text{ satisfying } ECR_{(1)}\left(\omega^*\right) = \min_{\omega} ECR_{(1)}(\omega)$$

Due to a complex nonlinear functional form of $ECR_{(1)}(\omega)$, an explicit solution for ω may not be tractable and thus the Nelder-Mead downhill simplex method is used to find ω^*, optimal length of warranty period. This method is known not to require the existence of derivatives of the objective function and is widely used in finding the solution for such a nonlinear function.

3.3 Optimal PM Period Under Renewable MRR Model

Although the frequent PM actions increase the warranty cost from the manufacturer's perspective, such actions may reduce the likelihood of the system failures. In this regard, it is necessary for the manufacturer to find an optimal PM strategy minimizing the total warranty cost during the warranty period.

In this subsection, we consider a renewable MRR warranty model, under which the periodic PM actions are taken by the manufacturer during the warranty period and obtain an optimal length of PM period between two successive PMs, which minimizes the expected total warranty cost from the manufacturer's perspective. Under the renewable MRR warranty, the warranty is renewed each time the broken system is replaced and thus the warranty is extended further proportionately to the number of replacements that occur while the warranty is in effect. Note that the warranty is terminated when no replacements occur in the renewed warranty period. The expected length of extended warranty under the renewable MRR warranty is given in Eq. (4).

To evaluate the expected total warranty cost incurred during the warranty period under the renewable MRR warranty with the PM actions, we consider replacement cost, minimal repair cost, PM cost and failure cost. In this study we assume that a certain portion of each replacement cost is charged to the user pro-rated to the usage duration of the system and all other warranty costs are at the expense of the manufacturer. Let ω denote the length of original warranty period and let $I_j (< \omega)$ denote the inter-replacement time interval elapsed between the $(j-1)st$ and the jth replacement of the system during the warranty period for $j = 1, 2, \cdots$. Then the length of extended warranty, denoted by ω_0, can be expressed as $\omega_0 = \sum_{j=1}^{N_R} I_j + \omega$. Let C_r, C_m, C_{pm} and C_f denote the *r.v.s* representing replacement cost, minimal repair cost, PM cost and failure cost incurred during the extended warranty period of length ω_0, respectively when the periodic PM period equals δ. Then the total warranty cost, denoted by $C(\omega, \delta)$, is obtained by summing these costs as $C(\omega, \delta) = C_r + C_m + C_{pm} + C_f$. Thus the expected total warranty cost, denoted by $EC_{(1)}(\omega, \delta)$, can be evaluated as $EC_{(1)}(\omega, \delta) = E(C_r) + E(C_m) + E\left(C_{pm}\right) + E\left(C_f\right)$.

Under the renewable MRR warranty model under study in this subsection, the failure rate is adjusted lower to improve the performance of the system by applying the Canfield's [4] failure rate reduction model at each PM during the warranty period. Let T and $h_{pm}(t)$ denote the failure time and failure rate of the system, respectively, when the periodic PMs are conducted during the warranty period. Then, by applying the inversion formula, the *pdf* and *cdf* of T can be obtained as functions of $h_{pm}(t)$ as follows.

$$f_{pm}(t) = h_{pm}(t) \cdot \exp\left\{ -\int_0^t h_{pm}(t)dt \right\} \text{ and } F_{pm}(t) = 1 - \exp\left\{ -\int_0^t h_{pm}(t)dt \right\},$$

where $h_{pm}(t)$ is the adjusted failure rate under Canfield's [4] model, which is given in Eq. (5).

We note that the sequence of random variables, $I_1, I_2, \cdots, I_{N_R}$, are *i.i.d.* with a mean of

$$E(I_1) = E(T | T < \omega, Y > r_0) = \frac{\int_0^\omega t \cdot f_{pm}(t)dt}{F_{pm}(\omega) \cdot \overline{G}(r_0)} = m(\omega, r_0). \tag{14}$$

Here, Y is the *r.v.* representing the length of repair time and we denote r_0 the repair time threshold. Thus, $m(\omega, r_0)$ can be interpreted as the mean length of life cycle of the new system that replaces the failed system during the warranty period of length ω. Note that the broken system is replaced by a new one only when the repair time exceeds the repair time threshold r_0.

Since the number of replacements occurred during the warranty period, denoted by N_R, has the following geometric distribution, the *pdf* of N_R has the following expression.

$$P(N_R = n) = \left\{ \overline{F}_{pm}(\omega) + F_{pm}(\omega) \cdot G(r_0) \right\} \cdot \left\{ F_{pm}(\omega) \cdot \overline{G}(r_0) \right\}^n,$$

$$n = 0, 1, 2, \cdots, \tag{15}$$

where G(·) is the *cdf* of repair time Y and $\overline{G}(\cdot) = 1 - G(\cdot)$.

It follows from Eq. (15) that the expected number of replacements occurred during the warranty period can be obtained as

$$\mathrm{E}(N_R) = \frac{F_{pm}(\omega) \cdot \overline{G}(r_0)}{1 - F_{pm}(\omega) \cdot \overline{G}(r_0)}. \tag{16}$$

Next, we derive the expected number of minimal repairs that would be carried out until the warranty expires. Note that the warranty expires when no replacements occur during the warranty period of the new system. Since the random variables $I_1, I_2, \cdots, I_{N_R}$ are *i.i.d.* with a mean of $m(\omega, r_0)$, we have $\mathrm{E}(I_1) = \mathrm{E}(I_2) = \cdots = \mathrm{E}\left(I_{N_R}\right) = m(\omega, r_0)$ by Eq. (14). Let $N_M\left(I_j\right)$ be the expected number of minimal repairs conducted in the interval $\left(0, I_j\right)$. Then, we have

$$N_M\left(I_j\right) = \int_0^{I_j} h_{pm}(t)dt. \tag{17}$$

By replacing I_j of integration in Eq. (17) with $m(\omega, r_0)$, we obtain the approximation of $\mathrm{E}N_M\left(I_j\right)$ as follows.

$$\mathrm{E}N_M\left(I_j\right) \cong \int_0^{m(\omega, r_0)} h_{pm}(t)dt. \tag{18}$$

Thus, the expected number of minimal repairs that would be carried out until the warranty expires, given $N_R = n$, can be evaluated as

$$\begin{aligned} \mathrm{E}(N_M | N_R = n) &= \sum_{j=1}^{n} E N_M\left(I_j\right) + \int_0^{\omega} h_{pm}(t)dt \\ &\cong \mathrm{n} \cdot \int_0^{m(\omega, r_0)} h_{pm}(t)dt + \int_0^{\omega} h_{pm}(t)dt. \end{aligned} \tag{19}$$

The second term in Eq. (19) is added to include the number of minimal repairs conducted during the renewed warranty period following the nth replacement, where no replacements occur.

The replacement cost during the warranty period under the renewable pro-rata MRR warranty can be expressed as a function of I_j, $j = 1, 2, \cdots$ as

$$\mathrm{C_r} = c_r \cdot \sum_{j=1}^{N_R} \left(1 - \frac{I_j}{\omega}\right).$$

Let T_j, $\mathrm{j} = 1, 2, \cdots$, denote the failure time of the system, which has been replaced at $I_1 + I_2 + \cdots + I_{j-1}$. Then, the expected replacement cost, conditioned on $N_R = n$, can be evaluated as

$$\begin{aligned}
\mathrm{E}(C_r|N_R = n) &= Ec_r \left\{ \left(\sum_{j=1}^{N_R} \left(1 - \frac{I_j}{\omega}\right) | N_R = n \right) \right\} \\
&= \sum_{j=1}^{N_R} c_r \left(1 - \frac{E\left(I_j\right)}{\omega} | N_R = n\right) \\
&= c_r \left\{ n - \frac{1}{\omega} \cdot \sum_{j=1}^{n} E\left(T_j | T_j \le \omega, Y_j \ge r_0\right) \right\} = c_r \left\{ n - \frac{n}{\omega} \cdot \frac{\int_0^{\omega} t \cdot f(t)dt}{F(\omega) \cdot \overline{G}(r_0)} \right\}
\end{aligned} \tag{20}$$

Here, Y_j is a *r.v.* representing the repair time corresponding to T_j. Using the fact that N_R has a geometric distribution as given in Eq. (15), we obtain the expected replacement cost by taking the expectation for $\mathrm{E}(C_r|N_R = n)$ of Eq. (20) with respect to N_R as follows,

$$\mathrm{E}(C_r) = E(\mathrm{E}(C_r|N_R = n)) = c_r \cdot \frac{F_{pm}(\omega) \cdot \overline{G}(r_0) - \frac{1}{\omega} \int_0^{\omega} t \cdot f_{pm}(t)dt}{1 - F_{pm}(\omega) \cdot \overline{G}(r_0)}. \tag{21}$$

Although the replacement cost is pro-rated to the user, we assume that the minimal repair cost, PM cost and failure cost are free of charge to the user. By multiplying the unit minimal repair cost, denoted by c_m, by $\mathrm{E}(N_M|N_R = n)$, given in Eq. (19), the conditional expectation on C_m, given that $N_R = n$, is obtained as

$$\begin{aligned}
E(C_m|N_R = n) &= c_m \cdot E(N_M|N_R = n) \\
&= c_m \cdot \left(\mathrm{n} \cdot \int_0^{m(\omega, r_0)} h_{pm}(t)dt + \int_0^{\omega} h_{pm}(t)dt \right).
\end{aligned} \tag{22}$$

By taking the expectation again on $E(C_m|N_R = n)$ of Eq. (22) with respect to N_R, the expectation for total minimal repair cost can be obtained as,

$$\mathrm{E}(C_m) = c_m \cdot \left(\frac{F_{pm}(\omega) \cdot \overline{G}(r_0)}{1 - F_{pm}(\omega) \cdot \overline{G}(r_0)} \int_0^{m(\omega, r_0)} h_{pm}(t)dt + \int_0^{\omega} h_{pm}(t)dt \right). \tag{23}$$

Let δ be the PM period between two successive PMs during the warranty period. Since the length of warranty period is extended proportionately to the number of replacements, it is assumed that the total number of PM actions conducted under our proposed renewable MRR warranty model is obtained by dividing the length of extended warranty period by the PM period. Thus, the PM cost during the extended warranty period can be evaluated as

$$C_{pm} = c_{pm} \cdot \left\lfloor \left(\sum_{j=1}^{N_R} I_j + \omega \right) / \delta \right\rfloor, \tag{24}$$

where $\lfloor x \rfloor$ denotes the integer part of number x.

By taking the double expectation on C_{pm} in Eq. (24), conditioned on $N_R = n$, we obtain $\mathrm{E}(C_{pm})$ as

$$\begin{aligned} \mathrm{E}(C_{pm}) &= E\left(E\left(C_{pm} | \sum_{j=1}^{N_R} I_j + \omega \right) \right) \\ &= c_{pm} \cdot \left\lfloor E\left(E\left(\sum_{j=1}^{N_R} I_j | N_R = n \right) \right) + \omega/\delta \right\rfloor \\ &= c_{pm} \cdot \left\lfloor \frac{\int_0^{\omega} t \cdot f_{pm}(t)dt}{1 - F_{pm}(\omega) \cdot \overline{G}(r_0)} + \omega/\delta \right\rfloor \end{aligned} \tag{25}$$

The failure cost incurs whenever the system failure occurs in the warranty period and the total failure cost is obtained by multiplying the unit failure cost by the total number of failures. Thus, we have

$$C_f = c_f \cdot (N_R + N_M).$$

By using the formula for $\mathrm{E}(N_M | N_R = n)$, given in Eq. (19), we can obtain the conditional expectation on C_f, given that $N_R = n$, as

$$\mathrm{E}(C_f | N_R = n) = c_f \cdot \left\{ \left(n + n \int_0^{m(\omega, r_0)} h_{pm}(t)dt \right) + \int_0^{\omega} h_{pm}(t)dt \right\}.$$

Thus, we have

$$\mathrm{E}(C_f) = E(E(C_f | N_R = n))$$

$$= c_f \cdot \left\{ \frac{F_{pm}(\omega) \cdot \overline{G}(r_0)}{1 - F_{pm}(\omega) \cdot \overline{G}(r_0)} \left(1 + \int_0^{m(\omega, r_0)} h_{pm}(t)dt \right) + \int_0^{\omega} h_{pm}(t)dt \right\} \tag{26}$$

Let $EC_{(1)}(\omega, \delta)$ denote the expected total warranty cost under the renewable pro-rata MRR warranty with the periodic PM actions from the manufacturer's perspective. Then, by adding each expected cost given in formulas (21, 23, 25 and 26), we obtain

$$EC_{(1)}(\omega, \delta) = E(C_r) + E(C_m) + E\left(C_{pm}\right) + E\left(C_f\right) \tag{27}$$

Park et al. [11] derive the formula given in Eq. (27). Utilizing Eq. (27) as an objective function to find an optimal value of δ, the optimization problem of minimizing the expected total cost under study is formulated as follows,

$$\text{Find } \delta^* \text{ satisfying } EC_{(1)}\left(\omega, \delta^*\right) = \min_{\delta} EC_{(1)}(\omega, \delta).$$

The objective function, given in Eq. (27), has a complex nonlinear functional form and its explicit solution for δ may not be obtainable analytically. Thus, the Nelder-Mead downhill simplex method is implemented to find δ^*, optimal length of inter-PM interval. You may refer to Park et al. [11] for more detailed discussions on the subject of this subsection.

4 Optimal Post-Warranty Maintenance Strategy from the user's Perspective

The manufacturer usually makes a certain type of warranty contract with the user at the sale of a system and the warranty provides the user a protection for the repairable system during the warranty period by assuring the operation of the system without failures. Recently, the system becomes more multi-functional and complex which makes the maintenance of the system more costly and difficult and thus, many users tend to prefer an extended warranty for a certain length of time following the expiration of the original basic warranty, especially for high-priced products or fragile kind of systems. To mention a few research works regarding the extended warranty, Lam and Lam [30] suggest an extended warranty model after a free replacement warranty is expired and Wu and Longhurst [31] investigate the operating cost of system warranted by basic and extended warranty policies with a non-renewing free replacement from the user's perspective. Bouguerra et al. [32] discuss the adoption of an extended warranty period for random failure systems with a free minimal repair warranty. Although the extended warranty may help the user to prolong the protection for the system longer, the user should take over the maintenance works for the system

once the warranty expires. After the expiration of warranty, the user is solely responsible for maintaining the system and is taking the risk of the system failures. In this regard, the post-warranty maintenance strategy becomes an increasingly important issue for the user in order not only to keep the system in the functioning conditions, but also to minimize the maintenance cost.

Sahin and Polatoglu [33] propose two types of post-warranty replacement policies subject to certain conditions and investigate the costs incurred during the life cycle of the system. Jung and Park [34] extend their replacement models by allowing the PM actions and propose the optimal maintenance policies in the post-warranty period. Jung et al. [35] propose an optimal replacement policy after the replacement warranty expires by developing a measure unifying the cost and downtime of the system. Various post-warranty maintenance policies have been studied incorporating many different types of warranty policies by Yeh et al. [36], Chen and Chien [37], Chien [38] and others. Yeo and Yuan [39] study a warranty model with the imperfect repair and extends the Yeh et al.'s [36] model. Jung et al. [40] define the life cycle anew from the customer's perspective and develop new optimal post-warranty maintenance policies under the renewable warranty policy. Later, Jung et al. [41] study an extended post-warranty maintenance model after the original two-phase warranty expires and determine an optimal length of maintenance period.

This section presents two optimal maintenance policies from the user's perspective after the warranty expires. During the warranty period, we adopt the renewable MRR warranty model which was first proposed in Park et al. [6]. Firstly, we consider a situation where the user maintains the system for a fixed length of time following the warranty period, which is referred to as a maintenance period throughout this section. During the maintenance period, only minimal repairs are conducted when the system failures occur and at the end of maintenance period the life cycle of the system ends and the system may be replaced by a new one. The second situation we consider in this section is as follows. The system undergoes the periodic PMs a pre-determined number of times following the expiration of warranty and conducts the minimal repair upon each inter-PM failures. At the time of last scheduled PM, the system is replaced by a new one ending the life cycle of the system. The optimization of post-warranty maintenance policy is obtained by finding the optimal values for the length of maintenance period and the length of inter-PM interval minimizing the expected cost rate per unit time in the life cycle of the system.

4.1 Optimization of Post-Warranty Maintenance Period Under Renewable MRR Warranty

During the warranty period, the maintenance of a system is mainly the responsibility of the manufacturer and whenever the system failure occurs, either the replacement or the minimal repair is carried out by the manufacturer. However, once the warranty expires the system maintenance is entirely the user's responsibility and thus, the user's

main concern would be to search for an optimal maintenance strategy in the post-warranty period. The criteria for the optimization is usually regarding the reliability of the system or the maintenance cost incurred in the life cycle of the system.

In this subsection, we consider a situation where a new system is purchased with the renewable MRR warranty with a fixed length of original warranty period and whenever a replacement occurs, the warranty is renewed with exactly the same warranty term as the original one. Note that the replacement takes place only when the repair time for the failed system exceeds the repair time threshold within the warranty period. Once the warranty expires, the maintenance of the system is solely the responsibility of the user during a fixed length of post-warranty maintenance period and only a minimal repair is provided whenever a failure occurs. At the end of maintenance period, the life cycle of the system ends. The main objective is to determine an optimal length of maintenance period minimizing the expected cost rate in the life cycle of the system. The life span of the system is defined as the length of time elapsed from the purchase of the system until the replacement of the system by a new one in the post-warranty maintenance period. Park et al. [6] propose the renewable MRR warranty model during the warranty period, which is discussed in Sect. 2 and incorporate such warranty model to obtain an optimal post-warranty maintenance policy.

Let C_r, C_m and C_f be the *r.v.*'s representing the replacement cost incurred during the warranty period and the costs for minimal repair and system failure during the life cycle of the system, for which the user is responsible, respectively. Then, the total maintenance cost charged to the user in the life cycle of the system would be equal to $E(C_r) + E(C_m) + E\left(C_f\right) + c_r$, where the last term c_r is added due to the assumption that the replacement is done at the user's expense when the life cycle of the system ends.

Under the renewable MRR warranty we consider in this subsection, the minimal repair cost and failure cost are assumed to be free of charge to the user in the warranty period. However, the replacement cost is charged to the user pro-rated in proportion to the usage of the system relative to the warranty period ω. Thus, the replacement cost charged to the user can be evaluated as a function of I_j's as follows.

$$\mathrm{C_r} = \sum_{j=1}^{N_R} c_r \cdot \frac{I_j}{\omega} \tag{28}$$

During the post-warranty maintenance period, only the minimal repair is conducted upon the system failure by the user with no repair time threshold. Since it is assumed that no minimal repair cost is charged to the user even during the warranty period, the user's minimal repair cost is obtained by multiplying the unit minimal repair cost, denoted by c_m, by the expected number of system failures during the post-warranty maintenance period. Assume that the failure process of the system follows the NHPP with intensity function of $\lambda(\cdot)$. Let N_{np} denote the number of failures during the post-warranty maintenance period. Then the *pdf* of N_{np} is given by

$$\mathrm{P}\big(N_{np}=n\big)=\frac{e^{-\int_{\omega}\omega+\delta_m\lambda(t)dt}\left\{\int_{\omega}^{\omega+\delta_m}\lambda(t)dt\right\}^{n}}{n!},$$

where δ_m denotes the length of post-warranty maintenance period. It follows that

$$\mathrm{E}\big(N_{np}\big)=\int_{\omega}^{\omega+\delta_m}\lambda(t)dt.$$

Since the failure cost during the post-warranty maintenance period incurs upon each failure, it is also obtained by multiplying the expected number of failures by the unit failure cost. Thus we have

$$C_m=c_m\cdot N_{np}\quad\text{and}\quad C_f=c_f\cdot N_{np} \tag{29}$$

By adding the costs obtained in Eqs. (28 and 29), the total maintenance cost charged to the user in the life cycle of the system can be expressed as

$$\mathrm{C}(\omega,\delta_m)=\sum_{j=1}^{N_R}c_r.\frac{I_j}{\omega}+N_{np}\cdot\big(c_m+c_f\big)+c_r.$$

Given $N_{np}=n$, the conditional expected total maintenance cost can be evaluated as

$$\begin{aligned}&\mathrm{E}\big(\mathrm{C}(\omega,\delta_m)|N_{np}=n\big)\\&=\frac{c_r}{\omega}\cdot\sum_{j=1}^{N_R}E\big(T_j|T_j\le\omega,Y_j\ge r_0\big)+E\big(N_{np}\big)\cdot\big(c_m+c_f\big)+c_r\\&=\mathrm{n}\cdot\frac{c_r}{\omega}\cdot\frac{\int_0^{\omega}t\cdot f(t)dt}{F(\omega)\cdot\overline{G}(r_0)}+\big(c_m+c_f\big)\cdot\int_{\omega}^{\omega+\delta_m}\lambda(t)dt+c_r\end{aligned} \tag{30}$$

By taking the expectation on $\mathrm{E}\big(\mathrm{C}(\omega,\delta_m)|N_{np}=n\big)$ of Eq. (30) with respect to N_{np}, the user's expected total maintenance cost in the life cycle of the system under the renewable MRR warranty model can be expressed as

$$\begin{aligned}EC_{(2)}(\omega,\delta_m)&=\mathrm{E}\big(\mathrm{E}\big(\mathrm{C}(\omega,\delta_m)|N_{np}=n\big)\big)\\&=\sum_{j=0}^{\infty}\big\{\mathrm{E}\big(\mathrm{C}(\omega,\delta_m)|N_{np}=j\big)\big\}\cdot\mathrm{P}\big(N_{np}=j\big)\\&=\sum_{j=0}^{\infty}\big\{\overline{F}(\omega)+F(\omega)\cdot\overline{G}(r_0)\big\}\cdot\big\{F(\omega)\cdot\overline{G}(r_0)\big\}^{j}\end{aligned}$$

$$\cdot \left\{ j \cdot \frac{c_r}{\omega} \cdot \frac{\int_0^{\omega} t \cdot f(t)dt}{F(\omega) \cdot \overline{G}(r_0)} + \left(c_m + c_f\right) \cdot \int_{\omega}^{\omega+\delta_m} \lambda(t)dt + c_r \right\}$$

$$= \frac{c_r}{\omega} \cdot \frac{\int_0^{\omega} t \cdot f(t)dt}{1 - F(\omega) \cdot \overline{G}(r_0)} + \left(c_m + c_f\right) \cdot \cdot \int_{\omega}^{\omega+\delta_m} \lambda(t)dt + c_r. \tag{31}$$

Under the renewable MRR warranty, the expected length of warranty period, denoted by $\mathrm{E}(L(\omega))$ is give in Eq. (4). Thus, the expected length of life cycle of the system is obtained as

$$\mathrm{E}(L(\omega)) + \delta_m = \frac{\int_0^{\omega} t \cdot f(t)dt}{1 - F(\omega) \cdot \overline{G}(r_0)} + (\omega + \delta_m). \tag{32}$$

By dividing $\mathrm{EC}(\omega, \delta_m)$, given in Eq. (31), by $\mathrm{E}(L(\omega)) + \delta_m$ of Eq. (32), we establish the expected cost rate per unit time in the life cycle of the system as

$$ECR_{(2)}(\omega, \delta_m) = \frac{\frac{c_r}{\omega} \cdot \frac{\int_0^{\omega} t \cdot f(t)dt}{1-F(\omega)\cdot\overline{G}(r_0)} + \left(c_m + c_f\right) \cdot \int_{\omega}^{\omega+\delta_m} \lambda(t)dt + c_r}{\frac{\int_0^{\omega} t \cdot f(t)dt}{1-F(\omega)\cdot\overline{G}(r_0)} + (\omega + \delta_m)}, \tag{33}$$

which can be used as an objective function to find an optimal length of post-warranty maintenance period. Letting r_0 become close to zero, $ECR_{(2)}(\omega, \delta_m)$ of Eq. (33) is reduced to

$$ECR_{(2)}(\omega, \delta_m)$$
$$= \frac{\frac{c_r}{\omega} \cdot \int_0^{\omega} t \cdot f(t)dt + \overline{F}(\omega) \cdot \left(c_m + c_f\right) \cdot \int_{\omega}^{\omega+\delta_m} \lambda(t)dt + \overline{F}(\omega) \cdot c_r}{\int_0^{\omega} t \cdot f(t)dt + \overline{F}(\omega) \cdot (\omega + \delta_m)}$$
$$\text{as } r_0 \to 0,$$

which is the same result as in Sahin and Polatoglu's [33] replacement model.

The optimization problem to find an optimal solution for δ_m, denoted by δ_m^*, can be formulated as

$$\text{Find } \delta_m^* \text{ satisfying } ECR_{(2)}\left(\omega, \delta_m^*\right) = \min_{\delta_m} ECR_{(2)}(\omega, \delta_m)$$

Again, the Nelder-Mead downhill simplex method is used to find δ_m^*, optimal length of post-warranty maintenance period from the user's perspective. Park et al. [6] present more detailed discussions regarding the renewable MRR warranty model and an optimal post-warranty maintenance policy, which is given in this chapter.

4.2 Optimization of Post-Warranty PM Period Under Renewable MRR Warranty

The post-warranty maintenance strategy adopted by the user gives significant influences not only on the performance of the system, but also on the maintenance cost charged to the user following the expiration of warranty. In this respect, a number of authors have been studying to search for an optimal post-warranty maintenance strategy incorporating different types of warranty policies adopted during the warranty period. Among them, Park et al. [6] introduce a renewable MRR warranty model to introduce an optimal post-warranty maintenance policy by determining the optimal length of maintenance period from the user's perspective.

Jung and Park [34] utilize Canfield's [4] failure reduction model for upgrading the failure rate at each PM to obtain the optimal number of PMs and the optimal length of inter-PM interval following the expiration of one-dimensional warranty by optimizing the post-warranty PM model. This subsection develops an optimal post-warranty PM policy minimizing the maintenance cost incurred during the post-warranty period from the user's perspective, which is an extension of Jung and Park's [34] maintenance model to the situation where a two-dimensional warranty is adopted depending on failure time and repair time simultaneously. The maintenance model we present in this subsection works as follows. During the warranty period, a renewable MRR model with a pre-determined repair time threshold is adopted and when the warranty expires, the user solely maintains the system by performing the PM actions $(L-1)$ times periodically. Then, at the next scheduled time of PM the system is replaced by a new and the life cycle of the system ends. For the system failures between two successive PMs, only a minimal repair is conducted.

The main goal of this subsection is to determine an optimal length of inter-PM interval during the post-warranty maintenance period, which minimizes the expected cost rate during the life cycle of the system. The life cycle of the system starts from the purchase of the new system and ends at the scheduled time of Lth PM during the post-warranty period. The following results are given in Park et al. [42].

Let C_r, C_m, C_f and C_{pm} denote the *r.v.'s* representing the replacement cost, minimal repair cost, failure cost and PM cost, respectively, incurred during the life cycle of the system from the customer's perspective. Then the total warranty cost, denoted by $\mathrm{C}(\omega, \delta_p)$, is obtained by summing these costs as $\mathrm{C}(\omega, \delta_p) = C_r + C_m + C_f + C_{pm}$, where δ_p denotes the length of inter-PM interval during the post-warranty period. Thus the expected total maintenance cost charged to the user during the life cycle can be evaluated as $\mathrm{EC}(\omega, \delta_p) = \mathrm{E}(C_r) + E(C_m) + E(C_f) + E(C_{pm}) + c_r$, where the last term c_r is added since it is assumed that the replacement occurs at the user's expense at end of the system's life cycle.

Under the renewable MRR warranty we consider in this subsection, the minimal repair cost is free of charge to the user and the replacement cost is prorated to the user as in subsection 4.1. However, the failure cost is entirely charged to the user in the life cycle of the system. Since the replacement cost incurs only during the warranty period from the user's perspective, we have

$$\mathrm{C_r} = \sum\nolimits_{j=1}^{N_R} c_r . \frac{I_j}{\omega},$$

which is exactly the same as Eq. (28). Thus, we obtain

$$\mathrm{E}(C_r) = c_r \cdot E(E(C_r|N_R = n)) = \frac{c_r}{\omega} \cdot \frac{\int_0^{\omega} t \cdot f(t)dt}{\left\{1 - F(\omega) \cdot \overline{G}(r_0)\right\}} \tag{34}$$

Let $H_m(\omega, \delta_p)$ denote the expected number of minimal repairs occurred during the post-warranty maintenance period where the PMs are conducted $(L-1)$ times periodically. At each PM, the failure rate is upgraded by applying Canfield's [4] failure rate reduction model, given in Eq. (5). Then by applying similar arguments as in Jung and Park [34], we obtain

$$H_m(\omega, \delta_p) = \left\{ \sum_{k=1}^{L-1} \sum_{i=1}^{k} \left\{ h\left((i-1)(\delta_p - \alpha) + (\delta_p + \omega)\right) - h\left(i(\delta_p - \alpha) + \omega\right) \right\} \right.$$
$$\left. \cdot \delta_p + h(t - k\alpha) + \sum_{k=0}^{L-1} \int_{k\delta_p + \omega}^{(k+1)\delta_p + \omega} h(t - k\alpha)dt \right\} \tag{35}$$

Since no minimal repair cost is charged to the user in the warranty period, the expected minimal repair cost can be evaluated by multiplying $H_m(\omega, \delta_p)$ of Eq. (35) by the unit minimal repair cost. Thus we have

$$E(C_m) = c_m \cdot H_m(\omega, \delta_p) \tag{36}$$

The periodic PMs are conducted $(L-1)$ times before the life cycle ends and thus the expected total PM cost can be obtained as

$$E(C_{pm}) = c_{pm} \cdot (L-1) \tag{37}$$

Let N_R denote the total number of system failures in the warranty period. Then the expected number of minimal repairs, given $N_R = n$, can be obtained by replacing h_{pm} in Eq. (5) by $\mathrm{h}(t)$ as

$$\mathrm{E}(N_M|N_R = n) \cong \mathrm{n} \cdot \int_0^{m(\omega, r_0)} h(t)dt + \int_0^{\omega} h(t)dt$$

The failure cost incurs at each failure of the system during the warranty period and thus, given that $N_R = n$, the expected failure cost during the warranty period can be expressed as

$$c_f \cdot \mathrm{E}(\mathrm{E}(N_R + N_M|N_R = n))$$

$$= c_f \cdot \left(\frac{F(\omega) \cdot \overline{G}(r_0)}{1 - F(\omega) \cdot \overline{G}(r_0)} \left(1 + \int_0^{m(\omega, r_0)} h(t)dt \right) + \int_0^{\omega} h(t)dt \right). \tag{38}$$

By adding $c_f \cdot H_m(\omega, \delta_p)$ to Eq. (38), the expected total failure cost incurred during the life cycle of the system is obtained as

$$\begin{aligned} \mathrm{E}(C_f) = c_f \cdot & \left(\frac{F(\omega) \cdot \overline{G}(r_0)}{1 - F(\omega) \cdot \overline{G}(r_0)} \left(1 + \int_0^{m(\omega, r_0)} h(t)dt \right) + \int_0^{\omega} h(t)dt \right) \\ & + c_f \cdot H_m(\omega, \delta_p), \end{aligned} \tag{39}$$

where $H_m(\omega, \delta_p)$ is given in Eq. (35).

Under the renewable MRR warranty, the expected length of warranty period, denoted by $\mathrm{E}(L(\omega))$, is give in Eq. (4). Thus, the expected length of life cycle of the system under study in this subsection is obtained as

$$\mathrm{E}(L(\omega)) + L \cdot \delta_p = \frac{\int_0^{\omega} t \cdot f(t)dt}{1 - F(\omega) \cdot \overline{G}(r_0)} + (\omega + \delta_p). \tag{40}$$

Let $EC_{(3)}(\omega, \delta_p) = E(C_r) + E(C_m) + E(C_{pm}) + E(C_f)$ denote the expected total maintenance cost under the renewable pro-rata MRR warranty with the periodic PM actions being conducted $(L - 1)$ times during the post-warranty maintenance period, where the formulas for $E(C_r)$, $E(C_m)$, $E(C_{pm})$ and $E(C_f)$ are given in Eqs. (34, 36, 37 and 39), respectively. Then the objective function for the optimization problem to optimize the length of inter-PM interval is the expected cost rate per unit time incurred during the life cycle of the system, denoted by $ECR_{(3)}(\omega, \delta_p)$. By dividing $EC_{(3)}(\omega, \delta_p)$ by $\mathrm{E}(L(\omega)) + L \cdot \delta_p$, we have

$$\begin{aligned} ECR_{(3)}(\omega, \delta_p) &= \frac{EC_{(3)}(\omega, \delta_p)}{\mathrm{E}(L(\omega)) + L \cdot \delta_p} \\ &= \frac{E(C_r) + E(C_m) + E(C_{pm}) + E(C_f)}{\frac{\int_0^{\omega} t \cdot f(t)dt}{1 - F(\omega) \cdot \overline{G}(r_0)} + (\omega + \delta_p)} \end{aligned} \tag{41}$$

The optimization problem to find an optimal solution for δ_p, denoted by δ_p^*, can be formulated as

$$\text{Find } \delta_p^* \text{ satisfying } ECR_{(3)}(\omega, \delta_p^*) = \min_{\delta_p} ECR_{(3)}(\omega, \delta_p)$$

The Nelder-Mead downhill simplex method is used to find δ_p^*, optimal length of inter-PM interval during the post-warranty maintenance period from the customer's

perspective. Park et al. [42] give more detailed discussions regarding the optimization problem given in this subsection.

5 Optimization of Two-Dimensional Warranty Based on Age and Usage with Lemon Law Application

This section presents an optimal two-dimensional warranty policy which considers both age and usage simultaneously to determine an optimal warranty length. The age of a system refers to its calendar age, which is measured as the time elapsed since the system was newly purchased, regardless of whether the system has been in operation or not, whereas the usage is counted as an actual operating time of the system. Obviously, the usage of a system is always shorter than the age except the situation where the system is operating continuously without interruption from the installment of the system. In many cases, the manufacturer suggests a two-dimensional warranty, which expires when the system arrives at the specified age or the specified usage. For instance, a typical warranty contract for new motor vehicle is a two-dimensional warranty consisting of an age limit (in years) and an odometer mileage limit (in mileage).

Another issue we consider in this section is the concept of refund which can be adopted instead of replacement upon the system failure. In many countries including USA, Europe and Asia, a regulation governing the failure and repair of the system named "lemon law" has been enacted to enforce the replacement or refund for the system which meets certain defective conditions, mostly for motor vehicles.

5.1 Description of Lemon Law

The lemon law [43, 44], which aims to provide more protection for the users who purchase defective motor vehicles, either new or second-hand, regulates that, when the failures occur more than a certain number of times within a certain time interval or an accumulated repair time exceed a pre-determined threshold, the manufacturer must replace the vehicle or make a refund to the user. The lemon law was mostly applicable to motor vehicles initially. However, its application becomes more widespread to other goods such as electrical appliances. Although the detailed conditions defining the defective system, which is referred to as a "lemon", are somewhat different depending on the country or the region, the system is determined to be a lemon in general by judging based on the number of failures and the accumulated repair time. For instance, in California, a system is declared "lemon" if the accumulated repair time exceeds 30 days or if the system fails more than a pre-specified failure number threshold. The lemon law being enacted in Korea since 2019 specifically regulates the conditions to determine the lemon and enforces the manufacturer to

make a refund to the user. Part of such conditions necessary to claim the refund are as follows. (i) the car has been owned by the original customer for less than a certain span, which is referred to as an age limit, (ii) the car has been driven for less than a certain distance, referred to as a usage limit, and (iii) the number of breakdown of either critical components or non-critical components exceed each failure number threshold.

In this section we discuss an optimal warranty policy applicable under lemon law by utilizing the similar techniques as being adopted in the renewable MRR warranty. The warranty conditions regulated under the lemon law is stochastically quite similar to the ones that is assumed in the renewable MRR warranty and Park et al. [45] suggest an optimal two-dimensional warranty strategy based on usage and ages by adopting some conditions regulated under the Korean lemon law and by utilizing the techniques presented in Sect. 2.

5.2 Prior Works Related to Lemon Law

The lemon law enforces the manufacturer to refund or replace the defective system with a new one if the system is declared to be a lemon and thus it affects the warranty policy significantly from the manufacturer's perspective. Since the lemon law became available in many countries, many researchers have worked on several aspects of lemon law which was originally enacted for the buyers of motor vehicles. The earlier works regarding the subject of lemon law are mostly concerned with the legal and economic aspects of the regulation brought about by protecting the buyers from purchasing the defective vehicles. Later, Park et al. [6] work on the renewable MRR warranty with a pre-determined repair time threshold in which the warranty benefit is shown to be similar to the obligation of the manufacturer enforced under the lemon law as far as the repair time is concerned. Since then Iskandar and Husniah [9] analyze the warranty cost incurred when implementing the lemon law for one-dimensional warranty in case the number of failures exceed a threshold. They also examine the refund and replacement cases, which is considered to be the first work dealing with the warranty policy under the lemon law. They extend the results to two-dimensional case later. Park and Park [12] investigate an optimal warranty strategy based on two factors, usage and age of the motor vehicle, applicable to the Korean lemon law and determine the optimal length of warranty period. Liu et al. [46] develop a maintenance model for warranty systems under lemon law attached to a rebate warranty policy with two limiting conditions. Husniah et al. [47, 48] extend Iskandar and Husniah's [9] results to the cases of multi-component system and second-hand system, respectively. Zhang et al. [49] study an optimal extension of two-dimensional warranty policy based on lemon laws to maximize the manufacturer's expected profit by balancing the trade-off between the warranty cost and the sales quantity.

It seems that there exist only a limited number of research works concerning the maintenance and warranty optimization with regard to the lemon law by very few authors in the literature so far. In the next subsection, we present the results of Park

et al. [45] which optimizes the two-dimensional warranty strategy based on usage and age by utilizing the techniques similar to those of the renewable MRR warranty.

5.3 *Optimization of Warranty and Refund Policy with Lemon Law Application*

The warranty model we consider in this subsection is similar to the renewable MRR warranty model discussed in Sect. 2. However, the major difference is that in this work the warranty is affected by two factors of age and usage of the system and the warranty period is partitioned into two parts, one for refund and the other for minimal repair. Consider the situation where the system is warranted as follows. At the sale of a new system, the manufacturer suggests a two-dimensional warranty policy based on the usage and age of the system, and the warranty expires when the system reaches the age-based or the usage-based warranty boundary, whichever comes first. In addition, each warranty period is partitioned into refund period and minimal repair period. The user can claim the refund only when the system is determined to be a lemon within the age-based or usage-based refund period. For the remaining warranty period, only a minimal repair is warranted upon each failure by the manufacturer. Once the warranty expires, the system is solely maintained by the user until its life cycle ends.

Figure 2 shows the relationship between the usage range and the age range in which the warranty and the refund are valid when the system is proved to be a lemon. In Fig. 2, W_L and U_L denote the warranty periods based on age and usage, respectively, whereas W_0 and U_0 denote the age limit and the usage limit for a valid refund period under the lemon law. The intervals $[0, W_0]$ and $[0, U_0]$ are referred to as refund periods throughout this subsection and the shaded region represents the minimal repair region.

In this study, we deal with the motor vehicle as a system regulated by the lemon law and the minimal repair cost is free of charge to the user and the refund cost is

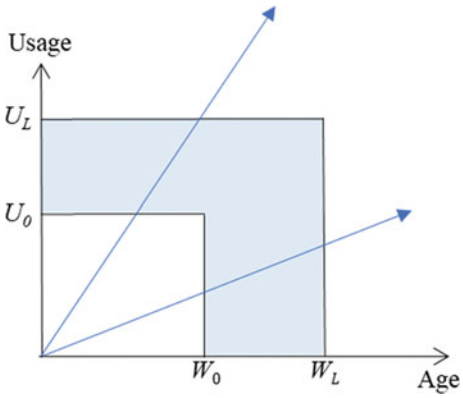


Fig. 2 Warranty regions with refund and minimal repair period

prorated to the user based on the usage of the vehicle when the vehicle is declared to be a lemon. When a refund is issued, the refund amount charged to the manufacturer, denoted by C_{re}, can be evaluated as

$$C_{re} = P_s \cdot \left(1 - \frac{U}{150,000}\right) + c_0, \tag{42}$$

where P_s, U and c_0 denote the original purchasing price, usage of the vehicle and other additional cost, respectively. When the system is recognized by a lemon, the manufacturer gives the user a refund, as given in Eq. (42) and the maintenance of the failed product is finished from the manufacturer's perspective. U is assumed to be distributed independently of the system failures in this work. Equation (42) is the equation proposed in the Korean Lemon Law to measure the refund amount and this equation reflects the fact that the refund amount decreases proportionately with increasing usage.

As for the structure of the system, we assume that the system has two components groups, critical components and non-critical components and we denote π_1 and π_2, respectively, the portions of critical and non-critical components in the system. For i = 1, 2, we let $f_i(\cdot)$, $F_i(\cdot)$, $\lambda_i(\cdot)$ and $\overline{F}_i(\cdot)$, respectively denote the *pdf*, *cdf*, intensity function, and reliability function for critical components ($i = 1$) and non-critical components ($i = 2$). One criterion to determine the defective vehicle as a lemon is the number of failures within the refund period. Thus we assume that a vehicle is declared to be a lemon when either the failure number for critical components reaches l_c or the failure number for non-critical component failures reaches l_n, whichever comes first, during the refund period.

Let L_c and L_n be *r.v's* representing the failure numbers for critical and non-critical component, respectively. Since the component failures follow the NHPP with the intensity functions of $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$, respectively, for the critical and non-critical components, the *pdf*'s for L_c and L_n can be expressed as

$$\begin{aligned} \mathrm{P}(L_c = n) &= \frac{\left[\int_0^W \lambda_1(t)dt\right]^n e^{-\int_0^w \lambda_1(t)dt}}{n!}, \\ \mathrm{P}(L_n = m) &= \frac{\left[\int_0^W \lambda_2(t)dt\right]^m e^{-\int_0^w \lambda_2(t)dt}}{m!} \end{aligned} \tag{43}$$

It follows that the expected failure numbers for critical and non-critical component, respectively, in an interval $[0, W]$ are given by

$$\mathrm{E}L_c = \int_0^W \lambda_1(t)dt, \quad \mathrm{E}L_n = \int_0^W \lambda_2(t)dt \tag{44}$$

In the next, we derive the expected warranty cost which can be used as an objective function for the optimization problem to find an optimal warranty length. For such a purpose, we adopt a univariate approach to make the two-dimensional process based on both usage and age to be a univariate model by defining the usage rate. Define the usage rate as

$$\mathrm{R} = \frac{U(t)}{t} \quad \text{for t} > 0,$$

where $U(t)$ denote the usage during the interval $[0, t]$. We assume that R is a *r.v.* with the *cdf* of $\mathrm{G}(r) = P(R \leq r)$. Given R = r, $U(t)$ can be expressed as $U(t) = r \cdot t$ and the conditional intensity function can be defined as $\lambda(t|r) = \psi(t, U(t))$, where $\psi(\cdot)$ is a non-decreasing function of t and $U(t)$. We assume that a minimal repair is carried out for the component failure with negligible repair time and for given R = r, the critical and non-critical component failures take place according to the NHPP with the conditional intensity functions $\lambda_1(t|r)$ and $\lambda_2(t|r)$, respectively. As an example of $\psi(\cdot)$, the following functional forms of conditional intensity functions are assumed for $\lambda_1(t|r)$ and $\lambda_2(t|r)$ in this study.

$$\lambda_1(t|r) = \theta_1 + \theta_2 r + \theta_3 T(t) + \theta_4 U(t), \lambda_2(t|r) = \theta_5 + \theta_6 r + \theta_7 T(t) + \theta_8 U(t) \tag{45}$$

where $\theta_i' s$, $i = 1, \cdots, 8$, are positive constants. These types of polynomial functions have been used in several related studies, such as Murthy et al. [50] and Su and Wang [24].

To formulate the expected warranty cost incurred by adopting the two-dimensional warranty under study, we define the age-based warranty period (W_L^r) and the age-based refund period (W_0^r) as follows.

$$W_L^r = \min\{W_L, U_L/r\} = \begin{cases} W_L, & if\ r \leq U_L/W_L \\ U_L/r, & if\ r > U_L/W_L, \end{cases} \tag{46}$$

and

$$W_0^r = \min\{W_0, U_0/r\} = \begin{cases} W_0, & if\ r \leq U_0/W_0 \\ U_0/r, & if\ r > U_0/W_0. \end{cases} \tag{47}$$

Here, the interval of $\left[0, W_0^r\right]$ is for the refund and within the interval of $\left[W_0^r, W_L^r\right]$, the manufacturer provides only minimal repairs upon the system failures.

Note that l_c and l_n denote the failure number thresholds for the critical and non-critical components, respectively, under the lemon law with $l_c < l_n$. This indicates that when either the failure number of critical component reaches l_c or the failure number of non-critical component reaches l_n within the interval $\left[0, W_0^r\right]$, whichever comes first, the manufacturer obligates to refund all or part of the original purchasing price to the customer under the lemon law, as defined in Eq. (42), and the life cycle of

the system ends. In case the numbers of critical and non-critical component failures are less than l_c and l_n within the interval $\left[0, W_0^r\right]$, each failure is only minimally repaired and the system is returned to the previous operating state. By utilizing the conditional intensity functions, denoted by $\lambda_1(t|r)$ and $\lambda_2(t|r)$, and by applying the NHPP failure processes for critical and non-critical failures, the *pdf*'s for L_c and L_n can be expressed as

$$\mathrm{P}(L_c = n) = \frac{\left[\int_0^{W_0^r} \lambda_1(t|r)dt\right]^n e^{-\int_0 W_0^r \lambda_1(t|r)dt}}{n!} \quad for\ n = 0, 1, \cdots \tag{48}$$

and

$$\mathrm{P}(L_n = m) = \frac{\left[\int_0^{W_0^r} \lambda_2(t|r)dt\right]^m e^{-\int_0 W_0^r \lambda_2(t|r)dt}}{m!} \quad for\ m = 0, 1, \cdots \tag{49}$$

Assuming independence of L_c and L_n, the probabilities of lemon caused by critical and non-critical component failures can be respectively calculated as $\mathrm{P}(L_c = l_c, L_n \leq l_n - 1) = \mathrm{P}(L_c = l_c)\mathrm{P}(L_n \leq l_n - 1)$ and $\mathrm{P}(L_n = l_n, L_c \leq l_c - 1) = \mathrm{P}(L_n = l_n)\mathrm{P}(L_c \leq l_c - 1)$. The probability that a lemon does not occur during the refund period can be calculated as $\mathrm{P}(L_c \leq l_c - 1, L_n \leq l_n - 1) = \mathrm{P}(L_c \leq l_c - 1)\mathrm{P}(L_n \leq l_n - 1)$.

To evaluate the expected warranty cost, we let c_c and c_n denote the unit minimal repair costs for critical component failures and non-critical component failures, respectively. If the system is declared to be a lemon, the refund is made to the user and the warranty is terminated from the manufacturer's perspective. Since the expected cost due to the critical component failures, denoted by $\mathrm{E}C_c$, can be evaluated by adding the expectations of repair cost for critical component failures, repair cost for non-critical component failures that occur before the lemon is declared and the refund cost. Thus we have

$$\begin{aligned}\mathrm{E}C_c = &\left[c_c\{l_c - 1\} + E(C_{re}) + \sum_{m=1}^{l_n - 1} m P\{L_n = m\}c_n\right] \\ &\mathrm{P}(L_c = l_c)\mathrm{P}(L_n \leq l_n - 1),\end{aligned} \tag{50}$$

where $E(C_{re})$ is obtained by replacing U with $\mathrm{E}(U)$ in Eq. (42). In a similar manner, the expected warranty cost due to the non-critical component failures, denoted by $\mathrm{E}C_n$, can be obtained as

$$\begin{aligned}\mathrm{E}C_n = &\left[c_n\{l_n - 1\} + E(C_{re}) + \sum_{n=1}^{l_c - 1} n P\{L_c = n\}c_c\right] \\ &\mathrm{P}(L_n = l_n)\mathrm{P}(L_c \leq l_c - 1)\end{aligned} \tag{51}$$

If no lemon is declared during the refund period, then only minimal repairs are provided during the interval of $\left[W_0^r, W_L^r\right]$. Thus the expected warranty cost when no lemon is declared, denoted by EC_{nl}, is evaluated as

$$\begin{aligned} \mathrm{EC}_{nl} = \Bigg[& c_c \sum_{n=1}^{l_c-1} n P\{L_c = n\} + c_n \sum_{m=1}^{l_n-1} m P\{L_n = m\} \\ & + c_c \int_{W_0^r}^{W_L^r} \lambda_1(t|r)dt + c_n \int_{W_0^r}^{W_L^r} \lambda_2(t|r)dt \Bigg] \cdot \\ & \mathrm{P}(L_c \le l_c - 1)\mathrm{P}(L_n \le l_n - 1). \end{aligned} \tag{52}$$

By adding Eqs. (50, 51 and 52), the expected total warranty cost of the system for given R = r, denoted by $C\left(W_L^r\right)$, during the warranty period can be evaluated as

$$C\left(W_L^r\right) = EC_c + EC_n + EC_{nl}. \tag{53}$$

Note that if l_c and l_n become equal to ∞, then the above approach would be used to obtain an optimal minimal repair policy without a refund.

By taking the expectation of W_L^r and $C\left(W_L^r\right)$, which are respectively given in Eqs. (46) and (53), with respect to r, the expected warranty length and the total expected warranty cost during the warranty period, denoted by $E(W_L)$ and $\mathrm{E}[C(W_L)]$, can be obtained as

$$E[W_L] = \int_{r_{\min}}^{r_{\max}} W_L^r dG(r) \quad and \quad E[C(W_L)] = \int_{r_{\min}}^{r_{\max}} C\left(W_L^r\right) dG(r), \tag{54}$$

respectively, where $\mathrm{G}(\cdot)$ is the *cdf* of *r.v.* R. The formula for $\mathrm{E}[C(W_L)]$, given in Eq. (54) evaluates the expected warranty cost from the manufacturer's perspective when both age and usage are considered for the two-dimensional policy when applying the lemon law conditions considered in this study.

By using Eq. (54), the expected cost rate during the warranty period is given by dividing the expected total warranty cost $\mathrm{E}[C(W_L)]$ by the expected warranty length $E(W_L)$ as

$$ECR_{(4)}(W_L) = \frac{E[C(W_L)]}{E(W_L)}. \tag{55}$$

Note that since we obtain the ECR during the warranty period, we use $E(W_L)$, not the expected length of time elapsed until the warranty is terminated by the occurrence of lemon within the warranty period, as a denominator in Eq. (55).

The optimal warranty length, denoted by W_L^*, can be obtained by minimizing the $ECR_{(4)}(W_L)$ as

$$ECR_{(4)}\left(W_L^*\right) = \min_{W_L} \frac{\mathrm{E}[C(W_L)]}{E(W_L)}.$$

For more complete discussions on the optimal warranty length under study in this subsection, you may refer to Park et al. [45].

6 Concluding Remarks

This chapter has reviewed several optimization problems to determine its optimal relevant decision variables minimizing the expected warranty costs during certain intervals, such as life cycle, warranty period or post-warranty maintenance period. The decision variables of our interest include the length of warranty period, inter-PM interval and length of post-warranty maintenance period. All of the warranty models presented in this chapter are based on the renewable minimal repair-replacement(MRR) warranty under which both failure time and repair time are considered at the same time upon the system failure. Furthermore, the warranty conditions under the MRR warranty is somewhat similar to the ones regulated under the lemon law which aims to protect the buyers of the defective motor vehicles and the warranty model applicable to the lemon law is also presented in this chapter.

As the system becomes more complex and multi-functional in recent years, the subject of optimization for the warranty and maintenance of the repairable system has emerged as very important issue to challenge in the field of reliability engineering and many authors have worked on this subject and proposed a number of optimal maintenance policies incorporating various types of warranty models. The MRR warranty is one of such warranty models that has been developed in 2013 and, although only a few maintenance policies utilizing the MRR warranty are reviewed in this chapter, many other research results concerning the optimization of warranty and maintenance strategy based on the MRR warranty have been proposed in the literature.

Recently, the optimal maintenance policy applicable under the lemon law attracts a lot of attention from many authors due to the fact that such policy affects the warranty cost from the manufacturer's perspective. Since the manufacturer is obligated to refund or replace the failed system when the defective system meets some conditions within a certain interval, the optimal maintenance policy in consideration of the lemon law would become an important issue to challenge in the future research works.

Acknowledgements The first author's work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2020R1F1A104823711). The second author's research was partially supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R111A1A01066212).

References

1. Blischke W (Ed) (2019) In: Warranty cost analysis. CRC Press
2. Blischke W (1995) Product warranty handbook. CRC Press
3. Park M, Pham H (2011) Maintenance and warranty concepts. In: Safety and risk modeling and its applications, Springer, London, pp 101–122
4. Canfield RV (1986) Cost optimization of periodic preventive maintenance. IEEE Trans Reliab 35(1):78–81
5. Kim DK, Lim JH, Park DH (2015) Optimal maintenance level for second-hand product with periodic inspection schedule. Appl Stoch Model Bus Ind 31(3):349–359
6. Park M, Jung KM, Park DH (2013) Optimal post-warranty maintenance policy with repair time threshold for minimal repair. Reliab Eng Syst Saf 111:147–153
7. Goldberg J, Paz L (1991) Locating emergency vehicle bases when service time depends on call location. Transp Sci 25(4):264–280
8. Snyder LV, Daskin MS (2005) Reliability models for facility location: the expected failure cost case. Transp Sci 39(3):400–416
9. Skandar BI, Husniah H (2016) Cost analysis of lemon law warranties. In: Proceedings of Asia-Pacific international symposium on advanced reliability and maintenance modeling. Seoul, Korea
10. Wang X, He K, He Z, Li L, Xie M (2019) Cost analysis of a piece-wise renewing free replacement warranty policy. Comput Ind Eng 135:1047–1062
11. Park M, Jung KM, Park DH (2017) Optimal maintenance strategy under renewable warranty with repair time threshold. Appl Math Model 43:498–508
12. Park M, Park DH (2018) Two-dimensional warranty policy for items with refund based on Korean lemon law. J Appl Reliab 18(4):349–355
13. Nakagawa T (2005) In: Maintenance theory of reliability. Springer Science & Business Media
14. Iskandar BP, Murthy DP, Jack N (2005) A new repair–replace strategy for items sold with a two-dimensional warranty. Comput Oper Res 32(3):669–682
15. Jung M, Bai DS (2007) Analysis of field data under two-dimensional warranty. Reliab Eng Syst Saf 92(2):135–143
16. Chen T, Popova E (2002) Maintenance policies with two-dimensional warranty. Reliab Eng Syst Saf 77(1):61–69
17. Ye ZS, Murthy DP, Xie M, Tang LC (2013) Optimal burn-in for repairable products sold with a two-dimensional warranty. IIE Trans 45(2):164–176
18. Shafiee M, Chukova S (2013) Maintenance models in warranty: a literature review. Eur J Oper Res 229(3):561–572
19. Shafiee M, Finkelstein M, Zuo MJ (2013) Optimal burn-in and preventive maintenance warranty strategies with time-dependent maintenance costs. IIE Trans 45(9):1024–1033
20. Huang YS, Gau WY, Ho JW (2015) Cost analysis of two-dimensional warranty for products with periodic preventive maintenance. Reliab Eng Syst Saf 134:51–58
21. Wang J, Zhou Z, Peng H (2017) Flexible decision models for a two-dimensional warranty policy with periodic preventive maintenance. Reliab Eng Syst Saf 162:14–27
22. Wang Y, Liu Z, Liu Y (2015) Optimal preventive maintenance strategy for repairable items under two-dimensional warranty. Reliab Eng Syst Saf 142:326–333
23. Huang YS, Huang CD, Ho JW (2017) A customized two-dimensional extended warranty with preventive maintenance. Eur J Oper Res 257(3):971–978
24. Su C, Wang X (2016) A two-stage preventive maintenance optimization model incorporating two-dimensional extended warranty. Reliab Eng Syst Saf 155:169–178
25. Lim JH, Kim DK, Park DH (2019) Maintenance optimization for second-hand products following periodic imperfect preventive maintenance warranty period. Appl Stoch Model Bus Ind 35(4):1077–1089
26. Khatab A, Diallo C, Sidibe IB (2017) Optimizing upgrade and imperfect preventive maintenance in failure-prone second-hand systems. J Manuf Syst 43:58–78

27. Yeh RH, Lo HC, Yu RY (2011) A study of maintenance policies for second-hand products. Comput Ind Eng 60(3):438–444
28. Su C, Wang X (2014) Optimizing upgrade level and preventive maintenance policy for second-hand products sold with warranty. Proc Instit Mech Eng Part O: J Risk and Reliab 228(5):518–528
29. Park M, Jung KM, Park DH (2016) Optimal warranty policies considering repair service and replacement service under the manufacturer's perspective. Ann Oper Res 244(1):117–132
30. Lam Y, Lam PKW (2001) An extended warranty policy with options open to consumers. Eur J Oper Res 131(3):514–529
31. Wu S, Longhurst P (2011) Optimising age-replacement and extended non-renewing warranty policies in lifecycle costing. Int J Prod Econ 130(2):262–267
32. Bouguerra S, Chelbi A, Rezg N (2012) A decision model for adopting an extended warranty under different maintenance policies. Int J Prod Econ 135(2):840–849
33. Sahin I, Polatoglu H (1996) Maintenance strategies following the expiration of warranty. IEEE Trans Reliab 45(2):220–228
34. Jung GM, Park DH (2003) Optimal maintenance policies during the post-warranty period. Reliab Eng Syst Saf 82(2):173–185
35. Jung KM, Han SS, Park DH (2008) Optimization of cost and downtime for replacement model following the expiration of warranty. Reliab Eng Syst Saf 93(7):995–1003
36. Yeh RH, Chen MY, Lin CY (2007) Optimal periodic replacement policy for repairable products under free-repair warranty. Eur J Oper Res 176(3):1678–1686
37. Chen JA, Chien YH (2007) Renewing warranty and preventive maintenance for products with failure penalty post-warranty. Qual Reliab Eng Int 23(1):107–121
38. Chien YH (2008) A general age-replacement model with minimal repair under renewing free-replacement warranty. Eur J Oper Res 186(3):1046–1058
39. Yeo WM, Yuan XM (2009) Optimal warranty policies for systems with imperfect repair. Eur J Oper Res 199(1):187–197
40. Jung KM, Park M, Park DH (2010) System maintenance cost dependent on life cycle under renewing warranty policy. Reliab Eng Syst Saf 95(7):816–821
41. Jung KM, Park M, Park DH (2015) Cost optimization model following extended renewing two-phase warranty. Comput Ind Eng 79:188–194
42. Park M, Jung KM, Park DH (2018) Optimization of periodic preventive maintenance policy following the expiration of two-dimensional warranty. Reliab Eng Syst Saf 170:1–9
43. "Lemon Law". Can be accessed at https://enwikipediaorg/wiki/Lemon_law
44. "California lemon law". Can be accessed at http://www/dmvorg/ca-california/automotive-law/lemon-lawphp
45. Park M, Iskandar BP, Park DH (2022) Optimization of two-dimensional warranty policy for repairable products based on age and usage, applicable to lemon law conditions. Appl Stoch Model Bus Ind 38(6):1144–1157
46. Liu P, Wang G, Su P (2021) Optimal maintenance strategies for warranty products with limited repair time and limited repair number. Reliab Eng Syst Safety 210:107554
47. Husniah H, Pasaribu US, Iskandar BP (2020) Warranty cost analysis for a multi-component product protected by lemon laws. In: IOP conference series: materials science and engineering, December, vol 1003(1). IOP Publishing, pp 012110
48. Husniah H, Wangsaputra R, Iskandar BP (2020) Cost analysis of lemon law warranties for used equipments. Int J Artif Intell 18(1):86–96
49. Zhang Z, He S, He Z, Dai A (2019) Two-dimensional warranty period optimization considering the trade-off between warranty cost and boosted demand. Comput Ind Eng 130:575–585
50. Murthy DP, Iskandar BP, Wilson RJ (1995) Two-dimensional failure-free warranty policies: two-dimensional point process models. Oper Res 43(2):356–366

Reliability and Opportunistic Maintenance of Floating Offshore Wind Turbines

He Li and C. Guedes Soares

Abstract This chapter reviews the state-of-the-art methods and procedures for the reliability and maintainability of floating offshore wind turbines. First, a new failure identification and critical failure determination schedule is introduced, according to which failure prevention actions are determined. Subsequently, failure rate and reliability analysis models are reviewed, in particular, the assessment of failure rates of floating offshore wind turbines based on the onshore counterpart data. Finally, an opportunistic maintenance model is described for better scheduling of the maintenance crew, allowing limited preventive maintenance after corrective maintenance. Overall, methods and procedures introduced in this chapter contribute to failure and risk management, reliability improvements, and maintenance strategy planning of floating offshore wind turbines and can apply to other complex systems.

Keywords Floating offshore wind turbine · Failure identification and prevention · Failure rate and reliability · Opportunistic maintenance

1 Introduction

The last decades have seen extensive developments in renewable energies, including, but not limited to, solar, hydro, nuclear, and wind energy [1]. Wind energy is booming and is taking over the market share of conventional fossil energy-based energy production structures [2]. Practical evidence is that with about 94 GW (Gigawatt) wind power capacity added in the single year of 2021, including more than 21 GW from offshore, the total wind power capacity around the world reached 837 GW, indicating a 12% year-over-year growth [3].

Wind energy is stepping into waters at the coasts as a consequence of satisfactory wind profiles and away from the dense crowd. Floating offshore wind is a relatively new concept which is also a promising technique proposed to deal with electricity

H. Li · C. Guedes Soares (✉)
Centre for Marine Technology and Ocean Engineering (CENTEC), Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
e-mail: c.guedes.soares@centec.tecnico.ulisboa.pt

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_13

generation in deeper (typically more than 50 m) waters [4]. Floating offshore wind technology holds good adaptability to all waters than other types of equipment. However, the longer distance to the coast and ports and the lower accessibility of floating wind farms challenge the operation and maintenance (O&M) efficiency of floating offshore wind turbines (FOWTs). The mentioned factors also pulled up the O&M cost of such devices. For instance, the O&M cost of floating offshore wind turbines can be 35% of the overall cost, at least 5% higher than bottom-fixed offshore wind turbines, knowing that the same cost of onshore devices is 20–25% [5, 6]. Hence, academia and industry turn to reliability and maintainability investigations to clarify failure properties of floating wind turbines, find preventive actions to prevent failures' occurrence, reduce failure effects, or conduct efficient maintenance to restart the failed devices.

Reliability and maintainability of FOWT include failure identification and prevention, failure data correction and reliability analysis, and maintenance strategy planning. To be specific:

- Failure identification and prevention are to determine failures that potentially happen to floating offshore wind turbines, identify their critical failure behaviours, and find preventive means so as to cut failure propagation chains of floating devices [7, 8].
- Failure data correction and reliability analysis evaluate the reliability and failure rate of FOWTs, components, and elements based on the collected failure data and the constructed analytic model that reflects failures of FOWTs [9, 10].
- Maintenance strategy planning is to decide the elements to be fixed, the maintenance actions, and the time to start the maintenance under restrictions of maintenance resources such as personnel, tools, vessels, and accessibility of floating wind farms [11, 12].

This chapter aims to review state-of-the-art methods and tools implemented in failure identification and prevention, failure data correction and reliability analysis, and opportunity maintenance strategy planning to support the reliability and opportunistic maintenance of FOWTs, see Fig. 1. The chapter's outcomes would contribute to a deep understanding of the failure mechanisms, failure properties and reliability issues, maintenance actions and requirements of FOWTs.

The rest of this chapter is organized as follows. Sections 2–4 review solutions on failure identification and prevention, failure data correction and reliability analysis, and opportunity maintenance strategy planning proposed and accepted by the ARCWIND project (http://www.arcwind.eu/), respectively. The conclusion is provided in Sect. 5.

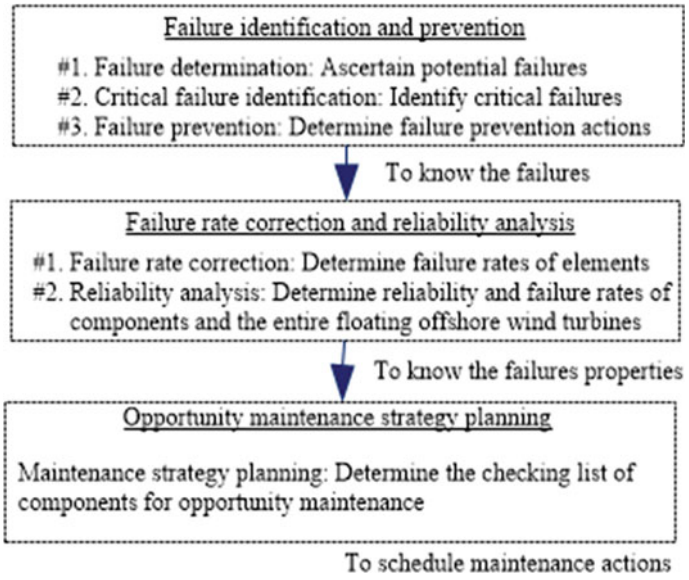


Fig. 1 Flowchart of failure, reliability, and maintainability investigations

2 Failure Identification and Prevention

Failure identification lists failures that may potentially happen to FOWTs during their 20–25 years of operation. Failure identification supports risk assesment, reliability analysis and improvement, availability accessment, and maintainance strategy palnning of FOWTs [13, 14]. It indicates the essential information to the mentioned subsequent studies, as these investigations are based on failures that already or potentially happen to FOWTs. To be specific: (1) risk analysis and control ascertain the most critical failures and design possible failure prevention actions [15]; (2) reliability is assessed based on the failure data statistics, representing a crucial performance index of systems [16]; (3) availability balances the normal working time and downtime resulting from failures, preventive maintenance, and windless conditions [17, 18]; (4) maintainability is a procedure of recovering/restarting floating offshore wind turbines from failures [19].

Failure prevention ascertains preventive actions such as design modifications, inspection and adjustment of the maintenance crew, resource preparation of supportive staff and suppliers, and so on. It prevents FOWTs from unwanted failures or a slight failure developing into a severe failure with unbearable consequences. The key to failure prevention is to determine failure behaviours and preventive actions to cut/impede the failure propagation chain so that critical failure causes would not happen or they would not give rise to critical consequences [8].

However, FOWTs (SPAR, Tension-Leg Planform, or Semi-Submersible structures) are relatively new concepts with very few installations [20–22]. Hence,

failure data from floating wind farms is still insufficient. Subsequently, reliability and maintainability related investigations of such devices are restricted by no-data conditions.

An applicable way to address data restrictions is to learn failure information from the onshore sector where the failure data is well accumulated, transforming failure data of floating devices from onshore ones. The following procedure was proposed to accomplish failure identification and prevention based on the transformed failure data, including potential failure determination, critical failure identification, and preventive action design.

2.1 Potential Failure Determination

Potential failure determination discovers failures that may happen to FOWTs during long-term operations [23]. As operational data of this kind of wind turbines is still unavailable, the potential failure determination relies on summarizing failures happened to onshore and bottom-fixed offshore equipment and expert judgments, see Fig. 2.

Failures of onshore and bottom-fixed offshore equipment provide a reference of FOWTs with the assistance of experts who determine: (1) if or not the onshore wind turbine failures would happen to float offshore facilities? (2) are there any additional failures that could happen to FOWT under the impact of harsh sea conditions?

The potential failure determination step consists of failure collection and expert adjustment application. Regarding Failure collection, 423 failures and their maintenance actions of 76 multi-MW (Megawatt) wind turbines in four wind farms were collected [24]. Overall, the observation consists of 1.44 million operation hours; the failure data of bottom-fixed offshore devicesd are collected from publications [5, 14, 25, 26]. With the information above, an initial failure sheet with 29 failure modes and 53 failure causes that happened in 15 components is summarized.

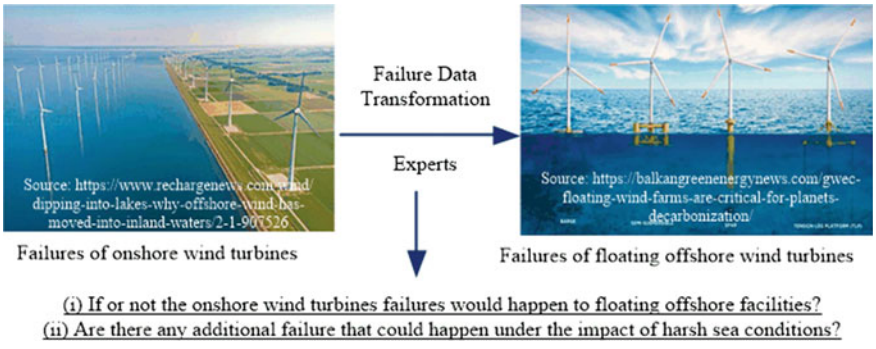


Fig. 2 Potential failure determination of FOWTs

Table 1 Specialists involved [6]

Experts	Duty	Working period/In the floating offshore sector
No. 1	System design (in company)	4 Years
No. 2	Components design (in company)	3 Years
No. 3	Quality engineer (in company)	4 Years
No. 4	Researcher (in university)	6 Years
No. 5	Chief technical officer (in company)	11 Years

Regarding expert adjustment, five experts were employed, see Table 1. The experts are mostly from wind companies (designers and maintenance crew), except for one from a university. The selection of the experts should be close to the industry and with different duties, guaranteeing the diversity of information sources and involving more knowledge and experiences. Another criterion of expert selection is the working period, the longer the better.

The experts' adjustment, see Fig. 3, determines the final failure sheet of FOWTs based on the onshore and bottom-fixed wind turbines' failures. The initial failure sheet was distributed to experts. They deleted unlikely failure items (in the initial failure sheet) or added new failure items, e.g., new failure causes to existing failure modes or new failures (component-failure mode-failure cause).

Failure Mode and Effect Analysis (FMEA) [27] is applied to critical failure identification. Accordingly, risk indices (severity, occurrence, and detection) are obtained by consulting with experts. Two documents should be prepared and distributed to experts: (1) the final failure list (Appendix A); and (2) rating guidance of indices, which can be three level, four level, ten level and other options.

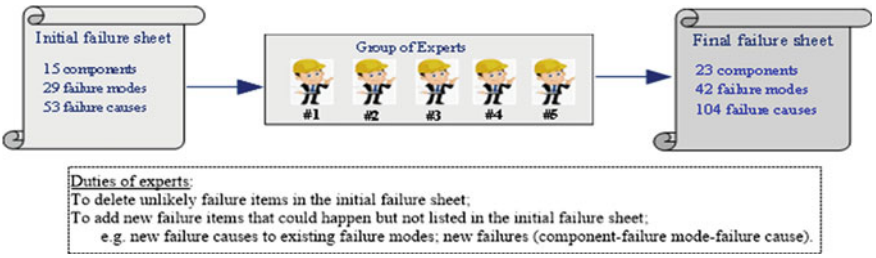


Fig. 3 Experts' adjustment

2.2 Critical Failure Identification

Critical failure identification determines critical failures of FOWTs, which is the connection between potential failure determination and failure prevention [28]. Critical failure identification selects the most critical failures of FOWTs and finds actions for failure prevention. In engineering, critical failure identification is mandatory as the maintenance resource and tools are limited to cover all failures. It is worth mentioning that not all potential failures would happen during the operation of floating offshore wind turbines, as the required inherent reliability and resistance have been designed and manufactured in the elements and components. But failures would still likely happen due to uncertainties in the material, manufacturing, supplier, installation, weather condition, operation, etc.

The main concern of the FMEA-based critical failure identification is that convincingly construct a Risk Priority Number (RPN) for each failure item. Critical failure items include critical components, dangerous failure modes, and risky failure causes. The purpose of failure identification is displayed in Fig. 4.

Conventional FMEAs are criticized for their weaknesses when applying to the critical failure identification of FOWTs, including [28]: (1) The subjective input data (severity, occurrence, and detection) makes RPNs subjective, without any physical meaning, and cannot be compared with others computed on different offshore wind turbines; (2) Assigning equal importance to experts and risk indices resulting in various failure items with the same RPN, and the hidden meaning of each could be completely different; (3) constructing decision-making indixes such as RPN based on subjective indices and ignored operational factors like failure cost; (4) independent failures assumptions. For convincingly constructing RPNs of failure items, several developments of FMEAs have been conducted, see Table 2, including:

1. Assign weights to indices [6]. A group of experts is employed. Unlike decision-making based on a single information source, critical failure identification becomes a group decision-making problem. Under this situation, assigning weight to evidence given by experts according to their working period (the years of working) in the floating offshore wind sector is required.
2. Assign weights to indices [8]. Severity, occurrence, and detection reflect different aspects of the risk of failure of FOWTs. However, the wind industry, especially its

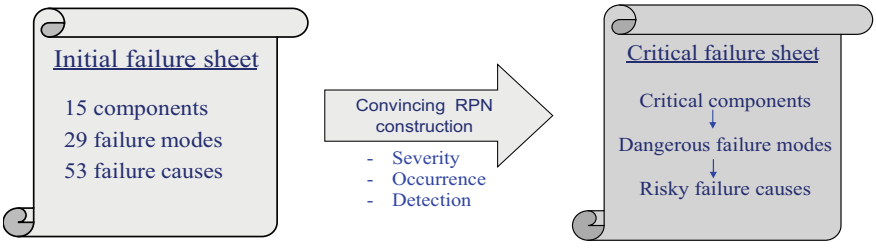


Fig. 4 The purpose of critical failure identification

Table 2 FMEA developments

Refs.	Development	Model property	Key index	Key parameter
[6]	Assign weights to experts	$\alpha_k = \frac{WP_k}{\sum_{k=1}^{i} WP_k}$	Weights of the expert k (α_k)	Working period (WP_k)
[8]	Assign weights to indices	$\boldsymbol{\omega_j} = \left[\frac{\sum_{j=1}^{3} \delta_{Sj}}{3} \quad \frac{\sum_{j=1}^{3} \delta_{Oj}}{3} \quad \frac{\sum_{j=1}^{3} \delta_{Dj}}{3} \right]^{T}$	Weights of the indices ($\boldsymbol{\omega_j}$)	Experts' judgments (δ_{Xj})
[28]	Model economic factors	$CRPN = S \times O \times D \times C$	Cost-and-risk Priority Number ($CRPN$):	Cost of failures (C)
[29]	Model failure correlations	$RPN(Final) = RPN(Fix.) + RPN(Var.) = S \times O \times D + S \times \Delta P \times D$	Failure correlation level (ΔP)	Bayesian network model

O&M, focuses more on the failure consequences (severity). Accordingly, highlighting severity in RPN calculation contributes to obtaining close-to-practice critical failures.

3. Model economic factors (costs of failures) [28]. FOWTs are economy-risk-key systems where one should consider both risks of failures and economic performance. Hence, modelling the cost of failures to reflect economic aspects during O&M would support robust decision-making in critical failure identification.
4. Model failure correlations [29]. FOWTs are complicated systems with correlated failures resulting from common cause failures (several failures sharing the same failure cause). Hence, in criticality modelling, the impact of a failure item on other failures should be considered.

Accordingly, critical failure items of FOWTs are identified with the assistance of the developed models in Table 2, including:

- Critical systems (contribute more than 50% RPN to the total): support structure (46%) and energy production system (33%);
- Critical components (contribute more than 50% RPN to the total): moorings (22%), generator (15%), gearbox (13%), and floating foundation (11%);
- Fourteen Critical failure modes (each contributes more than 2.4% RPN to the total): generator (3 failure modes, 3 FM), gearbox (1 FM), convertor (1 FM), transformer (1 FM), tower (2 FM), floating foundation (3 FM), and mooring (3 FM);
- Twenty-four Critical failure causes (each contributes more than 1% RPN to the total): generator (4 failure causes, 4 FC); gearbox (1 FC), converter (1 FC), transformer (1 FC), tower (5 FC), floating foundation (6 FC), and mooring (9 FC).

2.3 Failure Behaviours and Failure Preventions

Failure behaviour is a chain of failure initiation, propagation, and occurrence [28]. It indicates a failure cause results in a failure mode(s) of a component(s) and subsequently give rise to a failure of a system(s) until the malfunction of FOWTs. Fifteen failure behaviours are summarized according to the critical failure identified, see Fig. 5.

Failure prevention action is the preparation, measures, and activities for preventing critical failure from occurrence by cutting their propagation chains (failure behaviours). Accordingly, 18 failure prevention actions are found to avoid the critical failure behaviours identified, see Table 3.

3 Failure Data Correction and Reliability Analysis

3.1 Failure Rate Correction

Failure rates are inherent failure properties reflecting the failure likelihood of FOWT components under sea conditions [30]. Practically, failure rates are computed based on failure data statistics. The FOWT, however, is a new concept with limited installations and operation times. It resulted in a few failures collected and is sufficient for reliability, availability, and maintainability investigations. To this end, failure rate correction is mandatory to transform failure data of well-accumulated onshore wind turbines to floating offshore ones.

The failure occurrence in Sect. 2.2 is served as the benchmark of the failure rate correction, knowing that failure rate (objective evidence) and failure occurrence (subjective evidence) represent a similar physical index: failure likelihood [24]. The idea is to compute the differences between failure occurrences given by onshore (7 maintenance crew are employed) and floating offshore (5 experts in Sect. 2.1) experts and apply these differences in failure rate quantification. Accordingly, failure rates of elements of FOWTs can be calculated with the known failure rate of onshore wind turbines.

It is noted that at least two failure rate correction models are required to ascertain the failure rates of all items:

- For components with corresponding components in onshore wind turbines. The model is constructed based on the differences between failure likelihood of compoments equipted in both onshore wind turbines and FOWTs.
- For components without corresponding components in onshore wind turbines. The model is constructed based on the failure likelihood differences between similar components in FOWTs. Specifically, the failure rate of items in floating

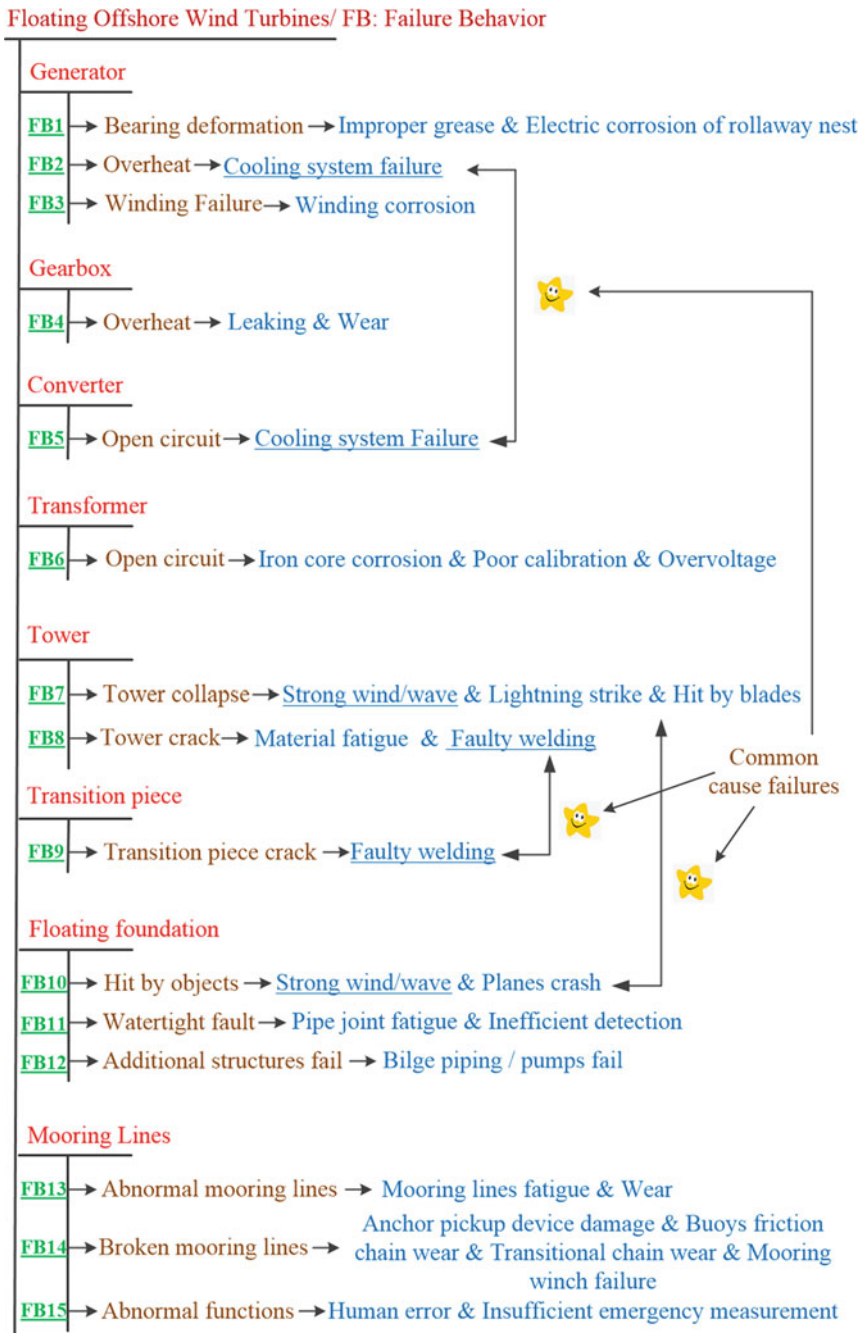


Fig. 5 Failure behaviours of FOWTs [8]

Table 3 Failure prevention action of FOWTs

System	Failure prevention actions
Energy-receiving system	• Blade: improve manufacturing processes and enhance blades test; • Blade: lightning prevention; • Hub and blade: improve installation processes for connection parts
Energy-producing system	• Main bearing: strengthen surfaces of bearing tracks; Welding quality inspections. • Generator: electric corrosion prevention; • Generator: upgrade cooling systems; • Gearbox: fatigue prevention of gears; • Generator and gearbox: high-quality lubrication
Energy-transforming system	• Converters and transformers: reliable cooling system; • Converters and transformers: prevent voltage/current overload offsetting
Auxiliary System	• Pitch: monitor pitch angles; • Yaw and pitch: high-quality hydraulic systems
Support structure	• Devices: improve mooring lines, anchor pickup device, transitional chain, mooring winch, fairlead, and accumulators; • Crew: operators and maintenance members training; • Tower: reinforce welding; • Structure failure: enhance pipe joint welding; platform monitoring; avoid transition pieces failures like corrosion and fatigue; • Environment factors: consider the impacts of strong wind/waves on support structures

foundations, transition pieces, and moorings are corrected according to the differences in failure occurrences between the mentioned components and mechanical components of onshore wind turbines like towers and nacelle.

It is concluded that the failure rate of FOWTs is 26% (based on the globalized model in [31]) or 28% (based on the localized model in [24]) higher than that of onshore wind turbines. The primary results are listed in Table 4 (at the subcomponent level) and Fig. 6 (at the component level).

3.2 Reliability Analysis

Reliability is the degree to which a FOWT operates as designed under a given time and working condition, based on which failure rate and mean time to failure (MTTF) can be computed [33]. Reliability analysis is a process of reliability computation. It infers the reliability of FOWTs and their components based on the failure rate corrected in Sect. 3.1.

Table 4 Failure rates and the corresponding correction factors of subcomponents [24]

Components	Subcomponents	Correction factor		Failure rates onshore	Failure rates corrected
		Amount	Rank		
Rotor	Blades	1.353	5	0.173	0.234
	Hub	1.224	7	0.069	0.085
	Main bearing	0.909	10	0.012	0.011
	Main shift	1.371	4	0.012	0.012
Generator	Generator	1.211	8	0.878	1.063
Gearbox	Gearbox	1.232	6	0.335	0.413
Electrical Facilities	Convener	1.531	3	0.693	1.060
	Transformer	1.555	2	0.042**	0.066
	Monitoring and SCADA	–	–	0.300	0.300*
	Weather Unit	–	–	0.127	0.127*
	Electronics and controller	1.688	1	0.730**	1.235
Pitch and yaw	Pitch system	0.906	11	0.416	0.377
	Yaw system	1.166	9	0.092	0.108
Cooling and hydraulic	Cooling system	–	–	0.809	0.809*
	Hydraulic	–	–	0.843	0.843*
Auxiliary	Crane	–	–	0.012	0.012*
	Climbing aid	–	–	0.012	0.012*
	Brake	–	–	0.035	0.035*
	Nacelle	–	–	0.104	0.104*

*: Failure rates without correction; **: Failure rates computed based on that reported in [8] and [32]; Failure rates in Failures/Year

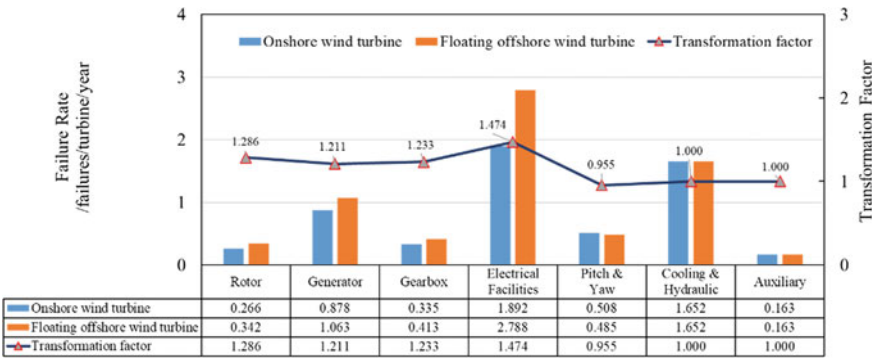


Fig. 6 Failure rates correction of components [24]

Bayesian networks (BNs) are applied to model the failure casualties of FOWTs, knowing that BNs can model correlated failures caused by common cause failures. Overall, two BNs are constructed for reliability computation, including:

- A hierarchical Bayesian network model for reliability analysis of support structures [34]. Failures of FOWTs are hierarchically structured, for instance, systems-components-failure causes, see Appendix A. A hierarchical Bayesian network model is constructed to model hierarchical failure causalities due to its better interpretability and capability to represent hierarchical systems.
- A Bayesian network model for reliability analysis of FOWTs [32]. A Bayesian network model completes the reliability analysis of entire FOWTs where correlated failures are considered.

It is concluded that a FOWT may fail 7.9 times per year with an MTTF of 1103 h [32]. Overall, the predicted failure rate and MTTF of FOWTs and their systems and components provide fundamental evidence for spare parts backup, logistic arrangement, maintenance resources preparation, and maintenance strategy planning of FOWTs and wind farms. Primary reliability analysis results of FOWTs are listed in Table 5 and Figs. 7 and 8.

Table 5 Reliability functions of FOWT systems (power function) [32]

Functions	Parameters			95% Confidence bounds			RMSE
	a	*b*	*c*	*a*	*b*	*c*	
SS	0.035	−5.723	4.0E−5	(0.035, 0.035)	(−6.564, −4.882)	(−9.1E−5, 1.7E−4)	1.519E−4
PS	0.314	−2.036	7.2E−3	(0.298, 0.330)	(−2.333, −1.739)	(−0.015, 6.1E−4)	5.952E−3
GB	2.889	−0.124	−2.153	(−0.614, 6.391)	(−0.299, 0.051)	(−5.676, 1.371)	1.744E−2
GE	0.411	−1.412	−0.025	(0.381, 0.440)	(−1.688, −1.315)	(−0.046, −0.004)	1.014E−2
AS	0.193	−2.450	−0.002	(0.186, 0.199)	(−2.709, −2.191)	(−0.005, 4.0E−4)	2.439E−3

SS Support Structure; *PS* Pitch System; *GB* Gearbox; *GE* Generator; *AS* Auxiliary System; *RMSE* Root Mean Square Deviation

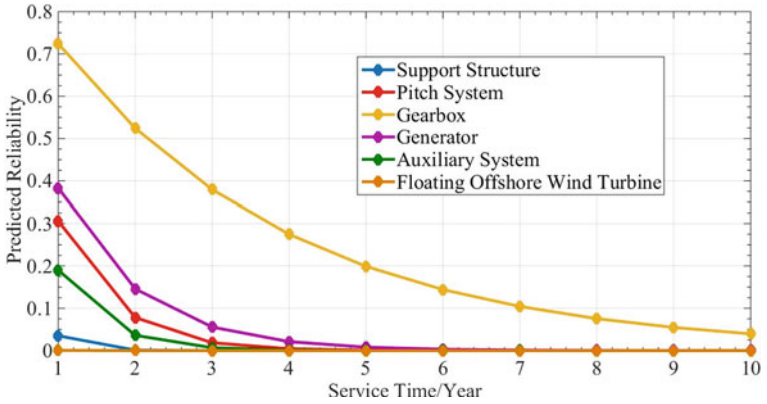


Fig. 7 Predicted reliabilities of FOWTs and their systems [34]

Fig. 8 Reliability analysis of systems of FOWTs [34]

4 Opportunistic Maintenance Strategy Planning

4.1 Maintenance Actions and Times

According to the analysis of 423 failures of the 76 multi-MW (Megawatt) wind turbines mentioned in Sect. 2.1, eight maintenance measures are implemented, see Table 6.

According to maintenance records, a maintenance procedure for a failure is divided into Reaction Time (RT), Traveling Time (TT), and Time to Repair (TR), see Fig. 9. The statistics of RT, TT, and TR of wind turbines are in Table 7.

Table 6 Maintenance measures

Maintenance measure	Action	Material asuumution	State of wind turbine	State of maintenance crew
Repair	Repairing	No	Working	None
Replace	Repairing	Yes	Working	None
Repaired and waiting for replacement	Repaired	Yes	Working	Waiting for replaceable parts
Checked and waiting for replacement (operating)	Checked without repair	No	Working	Waiting for replaceable parts
Checked and waiting for replacement (stopped)	Checked without repair	No	Stopped	Waiting for replaceable parts
Waiting for the supplier (operating)	Checked without repair	No	Working	Waiting for the maintenance crew of suppliers
Waiting for the supplier (stopped)	Checked without repair	No	None	Waiting for the maintenance crew of suppliers
Waiting for further instructions	The failure is unable to repair as (1) Unknow failure cause; (2) Lacking maintenance experience; (3) Minor failures with limited impact; (4) Beyond the authority of the maintenance crew; (5) Huge and expensive structural failures			

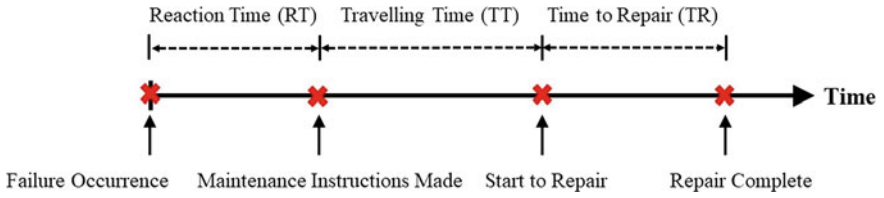


Fig. 9 RT, TT, and TR of wind turbines

4.2 Opportunistic Maintenance

Opportunistic Maintenance combines preventive and corrective maintenance, allowing to conduct limited preventative maintenance in the same visit for corrective maintenance [35]. The accessibility of floating offshore wind farms is low and restricted by weather windows, vessels, and other uncertain maintenance resources [36]. Hence, the crew should complete as many maintenance activities as possible in one visit to floating offshore wind farms. Opportunistic maintenance supports the maintenance efficiency improvement of FOWTs as it manages more preventive maintenance after the corrective maintenance has been completed, but no additional resources are required, including vessels, tools, labour, and so on.

Theoretically, opportunistic maintenance provides the maintenance crew with a checking list, according to which the maintenance crew conducts preventive maintenance within a limited period between planned corrective maintenance and return to port or motherboard. Following the potential failure identified in Sect. 2.1, the checking list prescribes clear information on failure and components to be checked. Figure 10 displays the logic of the checking list made for opportunistic maintenance of FOWTs.

An FMEA-BN model [35] is proposed to model the criticalities of FOWTs' failure items. The key idea of the model are as follows:

- Create a mirrored BN model according to the FMEA structure. The BN model reflects failure affiliations of the FMEA-based failure structure.
- Assign a risk index RI ($RI = S \times P \times D$, where S and D are in line with Sect. 2.2, P is updated by the BN model) to failure items of FOWTs to reflect the following operational states: operation conditions like the strong wind; failures impact of components on others; maintenance actions' impact on other components; The combinations of the above factors.
- The FMEA-BN model updates suggestions (checking list) according to (a) given operational state(s).

Four cases are examed to validate the effectiveness and applicability of the FMEA-BN model, including:

- Periodical preventive maintenance. It is to rank failure items to be inspected under the working condition;

Table 7 RT, TT, and MTTR of components by wind farms/hours [24]

Components	Subcomponents	Wind Farm #1			Wind Farm #2			Wind Farm #3			Wind Farm #4		
		RT	TT	MTTR	RT	TT	MTTR	RT	TT	MTTR	RT	TT	MTTR
Rotor	Blade	–	–	–	1236	–	0	–	–	–	0	2	4.5
	Hub	0	–	2	–	–	–	–	–	–	1	0	–
	Main bearing	0	7	0.6	–	–	–	–	–	–	–	–	–
Generator	Generator	0	3	2.4	5	10	4.3	8	0	9.8	280	6	4.4
Gearbox	Gearbox	1	5	2.3	0	–	–	3	12	4.7	1	–	2.6
Electrical Facilities	Converter	1	6	2	–	–	–	467	0	9.8	6	5	4.7
	Monitoring and SCADA	1	8	1	12	–	6	4	9	4.2	3	10	7
	Sensor	–	–	–	–	–	–	–	–	–	0	13	2
	Weather unit	–	–	–	13	–	9	0	4	5	3	0	7.8
Pitch and yaw	Pitch system	2	9	3.6	–	–	–	–	–	–	7	7	3.2
	Yaw system	0	0	4.2	0	–	0	0	–	–	–	–	–
Cooling and hydraulic	Cooling system	2	2	2.4	2	5	9	69	6.5	6.8	70	5	4.5
	Hydraulic	0	5	2.6	8	6	1.7	63	7	5.9	1	5	4.2
Auxiliary	Brake	–	–	–	14	12	9.5	–	–	–	–	13	2
	Nacelle	0	1	–	3	–	7	–	–	–	0	7	–
	Climbing aid	–	–	–	1	–	–	–		–	–	–	–
	Crane	–	–	–	–	–	–	–	–	–	0	0	–
Average		1	5	2.5	4	7	3.4	50	6	6.2	67	5	4.4

0: Less than one hour

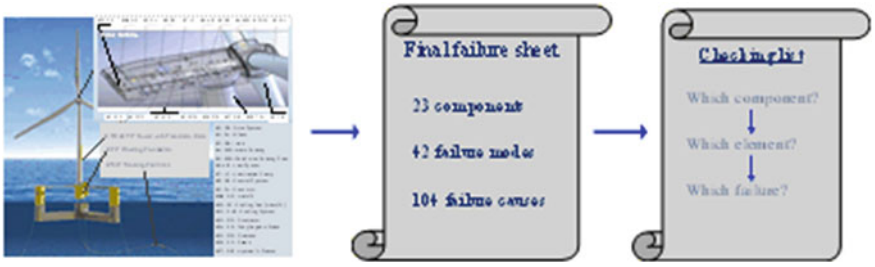


Fig. 10 Checking list for opportunistic maintenance

- Practical operation scenario. It is to determine failure items to be inspected under the situation of long-term unstable electricity output of wind turbines but no failures are reported by the SCADA;
- Specific maintenance was applied. It is to identify the failure items to be inspected after corrective maintenance of a component, such as generator, has been done;
- Dangerous weather conditions. It is to update the order of failure items to be inspected after dangerous weather conditions like storms.

Take the unstable electricity output as an example. Given the unstable electricity output state, the criticality of 11 failures is increased, but 6 decreased, see Table 8. The top 3 components in the checking list are the pitch subsystem, lubrication of generator bearings, and sensors placed on the generator.

Table 8 Criticalities of failure causes under the unstable electricity output state [35]

Failure causes	Criticalities changes		Failure causes	Criticalities changes	
	Amount	De/In-crease		Amount	De/In-crease
WT-BL-FM1-#1	10	↙	TT-TP-FM34-#58	13	↗
WT-MB-FM7-#7	26	↗	FF-FFE-FM35-#63	10	↙
WT-MSH-FM9-#9	11	↙	FF-FFE-FM37-#80	24	↗
WT-GE-FM11-#15	13	↗	MS-ML-FM38-#82	17	↗
WT-CV-FM20-#35	13	↗	MS-ML-FM38-#83	10	↗
WT-TR-FM22-#39	10	↗	MS-ML-FM38-#84	31	↗
WT-PS-FM23-#41	16	↗	MS-ML-FM39-#88	10	↙
WT-PS-FM25-#43	12	↗	MS-ML-FM39-#91	10	↙
TT-TO-FM31-#50	10	↙	↗/↙: Increased/Decreased		

5 Conclusion

This chapter reviewed the state-of-the-art methods and procedures proposed and conducted by the authors for reliability and maintainability issues of floating offshore wind turbines. It is concluded that: (1) a new failure identification and critical failure determination schedule is proposed according to which failure prevention actions are determined. The proposed program provides a new understanding of failures under the no-data situation; (2) failure rate correction and reliability analysis models are constructed and applied to assess the failure rates and reliability of floating offshore wind turbines and their components. The results provide urgently needed information, knowing that the same evidence is still insufficient for the floating offshore wind sector; (3) an opportunistic maintenance model is presented for a better arrangement of maintenance crew working period, which allows conducting limited preventive maintenance after corrective maintenance. The proposed model provides real-time opportunistic maintenance of floating offshore wind turbines. Overall, methods and procedures introduced in this chapter contribute to failure and risk management, reliability improvement, and maintenance strategy planning of floating offshore wind turbines and apply to other complex systems.

Ackowledgements This work contributes to the Strategic Research Plan of the Centre for Marine Technology and Ocean Engineering (CENTEC), which is financed by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e Tecnologia—FCT) under contract UIDB/UIDP/00134/2020.

Appendix: Failure modes and causes of the floating offshore wind turbine [8]

Failure mode level			Failure cause level	
Code	Failure modes	End effects	Code	Failure causes
WT-BL-FM1	Blades cracks	Wind turbine stop working	#1	Manufacturing error
WT-BL-FM2	Delamination	Wind turbine stop working	#2	Insufficient lighting protection
WT-BL-FM3	Gear teeth slip	Blades fail to attack wind properly	#3	Wear, fatigue, etc
WT-HB-FM4	Fracture in the shell	Rotor break	#4	Manufacturing error
WT-HB-FM5	Error in positioning	Blades break away from the hub	#5	Manufacturing error and/or fitting error
WT-MB-FM6	Bearing damage	Wind turbine stop working	#6	Wear, fatigue, etc

(continued)

(continued)

Failure mode level			Failure cause level	
Code	Failure modes	End effects	Code	Failure causes
WT-MB-FM7	Bearing vibration	Abnormal working condition	#7	Substandard lubrication
WT-MS-FM8	Cracks	Collapse of wind turbine	#8	Welding defects
WT-MS-FM9	Fracture	Collapse of wind turbine	#9	Fatigue
WT-GE-FM10	Bearing deformation	No, abnormal or unbalanced electricity generation	#10	Improper grease
			#11	Over tighten
			#12	Electric corrosion of rollaway nest
			#13	Shaft wear deformation
WT-GE-FM11	Overheat (GE)	Offshore wind turbine shutdown	#14	Turbine overload
			#15	Cooling system failure
			#16	Partial short circuit on stator winding
WT-GE-FM12	Winding failure	No, abnormal or unbalanced electricity generation	#17	Cable insulation failure
			#18	Connecting plug fall off
			#19	Interturn short circuit
			#20	Winding corrosion
WT-GB-FM13	Wear gears	Exceeded vibration or unstable electricity output	#21	Wear, fatigue
			#22	Dirty or lacking lubrication
WT-GB-FM14	Seized gears	No electricity output	#23	Sudden shock exceed limitation
WT-GB-FM15	Fractured gear teeth	Exceeded vibration or unstable electricity output	#24	Sudden shock exceed limitation
			#25	Fatigue
WT-GB-FM16	Wear bearing	Exceeded vibration	#26	Fatigue
WT-GB-FM17	Overheat (GB)	Offshore wind turbine shutdown	#27	Wear
			#28	Lubrication dried out
			#29	Leaking
WT-GB-FM18	Shift crack	Offshore wind turbine shutdown	#30	Fatigue
WT-CV-FM19	Short circuit	Converter shutdown	#31	Over heat
WT-CV-FM20	Open circuit	Disconnect to grid	#32	Load mutation

(continued)

(continued)

Failure mode level			Failure cause level	
Code	Failure modes	End effects	Code	Failure causes
			#33	Invert power input fault
			#34	Overload
			#35	Cooling system fault
WT-TR-FM21	Short circuit	Transformer shutdown	#36	Over heat
WT-TR-FM22	Open circuit	Disconnect to grid	#37	Constant overload
			#38	Iron core corrosion
			#39	Overcurrent
			#40	Overvoltage
WT-PS-FM23	Wrong pitch angle	Decrease of electricity output	#41	Poor calibration
WT-PS-FM24	Pitting Gears	Vibration increase	#42	Wear, fatigue
WT-PS-FM25	Misalignment bearings	Decrease of electricity output	#43	Wear, excessive vibration
WT-YS-FM26	Seizure bearings	Over heat	#44	Poor lubrication
WT-YS-FM27	Corrosions	Pitting of raceways	#45	Presence of corrosive substances
WT-YS-FM28	Hydraulic leakage	Rotor fails to stop	#46	Wear or degradation on hydraulic lines
WT-CE-FM29	Short circuit	Offshore wind turbine shutdown	#47	Moisture penetration
WT-CE-FM30	Open circuit	Offshore wind turbine shutdown	#48	Lightning strike
TT-TO-FM31	Tower collapse	Failure of whole facility and vast economic loses	#49	Strong wind/wave
			#50	Lightning strike
			#51	Hit by blades
			#52	Ice storm
			#53	Braking system failed
TT-TO-FM32	Abnormal vibration	Potential collapse	#54	Resonance
TT-TO-FM33	Crack	Potential collapse	#55	Faulty welding of Tower
			#56	Material fatigue
TT-TP-FM34	Transition piece crack	Potential collapse	#57	Material fatigue
			#58	Corrosion
			#59	Plastic deformation
			#60	Cyclic degradation
			#61	Strong wind/wave

(continued)

(continued)

Failure mode level			Failure cause level	
Code	Failure modes	End effects	Code	Failure causes
			#62	Faulty welding
FF-FF-FM35	Hit by dropped objects	Damage to the facility, vast economic loses	#63	Planes crash
			#64	Biological collision
			#65	Strong wind/wave
FF-FF-FM36	Watertight fault	Potential failure	#66	Inefficient detection
			#67	Pipe joint corrosion
			#68	Pipe joint weld defect
			#69	Pipe joint fatigue
			#70	Pillar damage
			#71	Excessive fouling of platform
FF-FF-FM37	Additional structures fail	Potential failure	#72	Navigation and work lights fail
			#73	Helicopter assistance equipment fail
			#74	Handrails corrosion
			#75	Ladders corrosion
			#76	Dynamic umbilical connection fail
			#77	Towing brackets/bollards fail
			#78	Vents fail
			#79	Bilge piping / pumps fail
			#80	Sensors for platform monitoring fail
			#81	Manholes fail
MS-ML-FM38	Abnormal mooring lines	Mooring line strength decrease or broken	#82	Mooring lines wear
			#83	Mooring lines fatigue
			#84	Mooring lines corrosion
			#85	Abnormal stress
			#86	Not effective maintenance
MS-ML-FM39	Mooring lines broken	Malfunction of the whole system, the facility cannot locate in water	#87	Transitional chain wear
			#88	Friction chain wear
			#89	Mooring winch failure

(continued)

(continued)

Failure mode level			Failure cause level	
Code	Failure modes	End effects	Code	Failure causes
			#90	Buoys friction chain wear
			#91	Anchor pickup device damage
			#92	Hydraulic motor failure
			#93	Accumulator failure
			#94	Over pressure
			#95	Connectors failure
			#96	Mooring interface structure failure
MS-FL-FM40	Fairlead failure	The anchor cannot be dropped and lift	#97	Fairlead corrosion
			#98	Fairlead fatigue
MS-AC-FM41	Anchor failure	Anchor failure	#99	Abnormal working conditions
			#100	Cyclic degradation
MS-RE-FM42	Abnormal functions	Anchoring accuracy decrease	#101	Poor operation environment
			#102	Insufficient emergency measurement
			#103	Human Error
			#104	Analysis and calculation fault

References

1. Olabi AG, Abdelkareem MA (2022) Renewable energy and climate change. Renew Sustain Energy Rev 158:112111
2. Díaz H, Serna J, Nieto J, Guedes Soares C (2022) Market needs, opportunities and barriers for the floating wind industry. J Marine Sci Eng 10(7):934
3. GWEC (2022) Global wind report 2022. https://gwec.net/global-wind-report-2022/. Accessed at 25 Nov 2022
4. Díaz H, Guedes Soares C (2020) Review of the current status, technology and future trends of offshore wind farms. Ocean Eng 209:107381
5. Sinha Y, Steel JA (2015) A progressive study into offshore wind farm maintenance optimisation using risk based failure analysis. Renew Sustain Energy Rev 42:735–742
6. Li H, Teixeira AP, Guedes Soares C (2022) An improved failure mode and effect analysis of floating offshore wind turbines. J Marine Sci Eng 10(11):1616
7. Ren Y, Venugopal V, Shi W (2022) Dynamic analysis of a multi-column TLP floating offshore wind turbine with tendon failure scenarios. Ocean Eng 245:110472

8. Li H, Díaz H, Guedes Soares C (2021) A failure analysis of floating offshore wind turbines using AHP-FMEA methodology. Ocean Eng 234:109261
9. Bhardwaj U, Teixeira AP, Guedes Soares C (2019) Reliability prediction of an offshore wind turbine gearbox. Renew Energy 141:693–706
10. Shafiee M (2023) Failure analysis of spar buoy floating offshore wind turbine systems. Innov Infrastruct Solut 8(1):1–19
11. McMorland J, Collu M, McMillan D, Carroll J (2022) Operation and maintenance for floating wind turbines: a review. Renew Sustain Energy Rev 163:112499
12. Kang J, Guedes Soares C (2020) An opportunistic maintenance policy for offshore wind farms. Ocean Eng 216:108075
13. Li H, Guedes Soares C (2021) A FMEA for a floating offshore wind turbine considering costs of failures. In: Guedes Soares C (ed) Developments in maritime technology and engineering. CRC Press, Singapore, pp 239–244
14. Kang J, Sun L, Guedes Soares C (2019) Fault tree analysis of floating offshore wind turbines. Renew Energy 133:1455–1467
15. Hou G, Xu K, Lian J (2022) A review on recent risk assessment methodologies of offshore wind turbine foundations. Ocean Eng 264:112469
16. Carroll J, McDonald A, McMillan D (2016) Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. Wind Energy 19(6):1107–1119
17. Cevasco D, Koukoura S, Kolios AJ (2021) Reliability, availability, maintainability data review for the identification of trends in offshore wind energy applications. Renew Sustain Energy Rev 136:110414
18. Lotovskyi E, Teixeira AP, Guedes Soares C (2022) Availability analysis of an offshore wind turbine subjected to age-based preventive maintenance by petri nets. J Marine Sci Eng 10(7):1000
19. Hameed Z, Vatn J, Heggset J (2011) Challenges in the reliability and maintainability data collection for offshore wind turbines. Renew Energy 36(8):2154–2165
20. Bagbanci H, Karmakar D, Guedes Soares C (2012) Review of offshore floating wind turbines concepts. Maritime engineering and technology, pp 553–562
21. Mas-Soler J, Uzunoglu E, Bulian G, Guedes Soares C, Souto-Iglesias A (2021) An experimental study on transporting a free-float capable tension leg platform for a 10 MW wind turbine in waves. Renew Energy 179:2158–2173
22. Uzunoglu E, Guedes Soares C (2020) Hydrodynamic design of a free-float capable tension leg platform for a 10 MW wind turbine. Ocean Eng 197:106888
23. Li H, Diaz H, Guedes Soares C (2021) A developed failure mode and effect analysis for floating offshore wind turbine support structures. Renew Energy 164:133–145
24. Li H, Peng W, Huang CG, Guedes Soares C (2022) Failure rate assessment for onshore and floating offshore wind turbines. J Mar Sci Eng 10(12):1965
25. Scheu MN, Tremps L, Smolka U, Kolios A, Brennan F (2019) A systematic failure mode effects and criticality analysis for offshore wind turbine systems towards integrated condition based maintenance strategies. Ocean Eng 176:118–133
26. Kang J, Sun L, Sun H, Wu C (2017) Risk assessment of floating offshore wind turbine based on correlation-FMEA. Ocean Eng 129:382–388
27. Huang J, You JX, Liu HC, Song MS (2020) Failure mode and effect analysis improvement: a systematic literature review and future research agenda. Reliab Eng Syst Saf 199:106885
28. Li H, Teixeira AP, Guedes Soares C (2020) A two-stage failure mode and effect analysis of offshore wind turbines. Renewable Energy 162:1438–1461
29. Sun Y, Li H, Sun L, Guedes Soares C. Failure analysis of floating offshore wind turbines with correlated failures. Submitted for publication
30. Cândido CJ, Santos SP (2015) Strategy implementation: What is the failure rate? J Manag Organ 21(2):237–262
31. Li H, Guedes Soares C (2022) Assessment of failure rates and reliability of floating offshore wind turbines. Reliab Eng Syst Saf 228:108777

32. Li H, Guedes Soares C, Huang HZ (2020) Reliability analysis of a floating offshore wind turbine using Bayesian networks. Ocean Eng 217:107827
33. Xu Z, Saleh JH (2021) Machine learning for reliability engineering and safety applications: review of current status and future opportunities. Reliab Eng Syst Saf 211:107530
34. Li H, Guedes Soares C (2019) Reliability analysis of floating offshore wind turbines support structure using hierarchical Bayesian network. In: Beer M, Zio E (eds) Proceedings of the 29th European safety and reliability conference. Research Publishing Services, Singapore, pp 2489–2495
35. Li H, Huang CG, Guedes Soares C (2022) A real-time inspection and opportunistic maintenance strategies for floating offshore wind turbines. Ocean Eng 256:111433
36. Guanche R, Martini M, Jurado A, Losada IJ (2016) Walk-to-work accessibility assessment for floating offshore wind turbines. Ocean Eng 116:216–225

A Summary of Inspection Policies of One Shot Systems

Qian Qian Zhao, Ha Won Kim, and Won Young Yun

Abstract In this chapter, we consider one shot systems, long-term repairable storage systems, which are in storage and can be used at an unknown time point once, whose failure only can be detected by inspection. Due to the system failure can incur a loss of life and economic damage, inspections and maintenance should be carried out to maintain a high level of storage reliability. However, since these inspections are usually costly, inspection times should be optimized to achieve a balance between undetected failure costs and inspection costs. Therefore, it is necessary to suggest appropriate optimization criteria and inspection policies according to the system structure and function characteristics of one shot system. In recent years, performance evaluation and inspection optimization problems have attracted many researchers' attention. This study summarizes the existing literatures related to the reliability and inspection optimization models of one shot systems. Firstly, this paper reviews the recent advances in storage reliability modeling for evaluating the performance of one shot systems. On this basis, the inspection optimization models of one shot systems with various structures are established and the key ideas of optimization methods in each optimization inspection problem are summarized. In summary, this contribution provides a survey on optimization methods for the inspection policy of one shot systems, with emphasis on the optimization methods under the different scales of systems, such as single-unit and multi-unit, as the target system. In addition, a qualitative comparison is performed to provide some general guidelines for the range of applicability of the approaches discussed in this contribution.

Keywords One shot system · Inspection policy · Storage availability · Methodologies

Q. Q. Zhao
School of Management, Zhengzhou University, Henan 450001, China

H. W. Kim
Rolling Stock Department RAMS Team, DAWONSYS, 41-20 Burim-ro170, Dongan-Gu, Anyang 14055, South Korea

W. Y. Yun (✉)
Majoring in Industrial Data Science and Engineering, Department of Industrial Engineering, Pusan National University, Busan 46241, South Korea
e-mail: wonyun@pusan.ac.kr

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_14

1 Introduction

A 'one shot' system is defined as a system that only can be used once, after use; the system is destroyed or must undergo an extensive rebuild. The system typically speeds its life in dormant storage readiness. Generally, storage reliability is used to evaluate the ability of a dormant system to keep its required function and is defined as the probability that the system can perform the expected function on demand after a certain storage time. The system storage reliability gradually declines with time, because some of its units are affected by various environmental pressures and then degrade with time, such as oxidation, corrosion, rust, and aging, and then suddenly fail at a random time, but the existing failure cannot be detected. The failure states of the one shot system can only be identified when it is inspected or used. Therefore, to prevent it from failing to perform its function when needed, we need to inspect periodically whether it is in a working state and maintain it as required.

For these repairable systems, system availability is more appropriate as the system performance measure. Many others have also researched storage availability models: See Cui et al. [1], Kim and Yun [2], Kitagawa et al. [3, 4] and Zhao and Yun [5]. Among them, this study mainly reviewed main models of the storage availability and inspection policies of single-unit and multi-unit one shot systems. During storage, the availability of the one shot system can be improved by timely inspection and maintenance of failures. The storage environment of this kind of system is strict, and changes in the environment could accelerate the degradation of the units. Excessive removal of the system from the original environment may reduce storage reliability and incur high inspection and maintenance costs. Therefore, the inspection interval for one shot systems should be optimized to achieve a balance between the system availability and maintenance costs.

Most inspection optimization models for the one shot system considered instantaneous and mean availabilities [3, 4], and overhaul time [6], which are related to system performance, as one of the optimization criteria. The expected cost rate [3, 4, 6–10] and the expected total cost until replacement or detection of failure [1, 11–13] are taken as another cost-related optimization criterion. Wolde and Ghobbar [14] considered reliability, availability, and cost as optimization criteria in their inspection optimization model. Researches on the inspection optimization model based on periodic inspection has been done for decades, and various inspection policies have been proposed. Refer Ito and Nakagawa [6, 8, 11, 12], Cho and Lee [10], Kitagawa et al. [4, 9]. For an aging system with increasing failure rate, the inspection policy with periodic interval is not optimal. It would be reasonable to inspect less frequently when the system is of its early age and to inspect more frequently as they deteriorate. The inspection interval should be shortened with time. Kitagawa et al. [3] and Nakagawa et al. [13] considered a non-periodic(sequential) policy in their optimization model assuming that units in one shot systems have following the Weibull distribution with an increasing failure rate.

The optimization of non-periodic(sequential) inspection is always a challenging problem in reliability area. The determination of each inspection time point is related

to the maintenance history and maintenance model of each unit after the inspection. Meta-heuristics have been developed for inspection scheduling problems under the non-periodic inspection policy. For example, Golmakani and Moakedi [15] proposed an A* search algorithm to determine an optimal non-periodic inspection scheme for a multi-unit repair system, which is based on a branch-and-bound method. Kitagawa et al. [3] used a simulated annealing (SA) algorithm to obtain optimal inspection intervals and the number of minimal repairs until replacement under an aperiodic inspection policy. A genetic algorithm (GA) with a heuristic method has been applied for the inspection optimization models of multi-unit systems [16, 17]. Zhao and Yun [18] designed a hybrid estimation of the distribution algorithm (EDA) to find the optimal inspection time points to minimize the life cycle cost and satisfy the interval availability required.

For the one shot system, in addition to inspection policies, optimization problems such as operational use (launching order) [19], and functional redundancy allocation [20] were also studied, which are not included in this review study. Although a number of inspection optimization models for various systems in reliability area have been published, there are a limited number of inspection models for one shot systems. Therefore, this study mainly reviews the existing inspection optimization models and methodology for one shot systems, intending to serve as guidance to both engineers and researchers so that decision-makers in reliability area of one shot systems can find the optimal inspection strategy of the one shot system. To summarize, in the area of reliability and cost optimization of one shot systems, the problem being actively studied is to determine the inspection intervals when a given target value is satisfied and the optimization problems to achieve this goal are classified into:

- Calculating the storage reliability/availability of a one shot system (considering inspection equipment errors).
- Determination of periodic/sequential inspection interval of a one shot system with repair.
- Determining the optimal inspection interval of a one shot system consisting of multiple units.
- Determination of the optimal inspection time points and preventive maintenance threshold of a one shot system that satisfies the target interval availability.

This chapter is organized as follows. In Sect. 2, we introduce the one shot systems and summarized the characteristics of the system operation environment [21], system structures [22], and unit failure modes. Section 3 discusses the performance measures for one shot systems, including instantaneous availability and interval availability models. Storage availability models are established for a single-unit one-shot system and a one shot system with multi-units. Then, several optimum inspection models for one shot systems under different inspection and maintenance policies are reviewed in Sect. 4. Section 5 concludes the paper with directions for further research.

2 One Shot Systems

One shot system is a storage system that remains inactive (non-operating mode) in a particular environment for a long period of time when no missions need to be performed [23]. For example, fire extinguishers, missiles, and airbags are typical one shot systems. As a multi-unit system, such as the mid-range infantry-guided missile (MRIGM) system, the one shot system consists of missile-like one shot devices and portable or vehicle-mounted facility-like support equipment, as shown in Fig. 1.

In general, units with a series structure in the one shot devices are usually divided into two types of independent units. Type 1 units in missiles are electronic devices (e.g., coolant pump, guidance unit) that fail at random times during storage period according to general failure distributions. In the existing literature, exponential and Weibull distributions are commonly used for describing the time to failures of type 1 units. The failed units can be identified at inspection. Corrective maintenance (repair/replacement) is done to restore their function. If the failure time of the type 1 unit follows an exponential distribution, we need to detect the failures of the units but it is not effective to maintain the units preventively. If the failure time of the

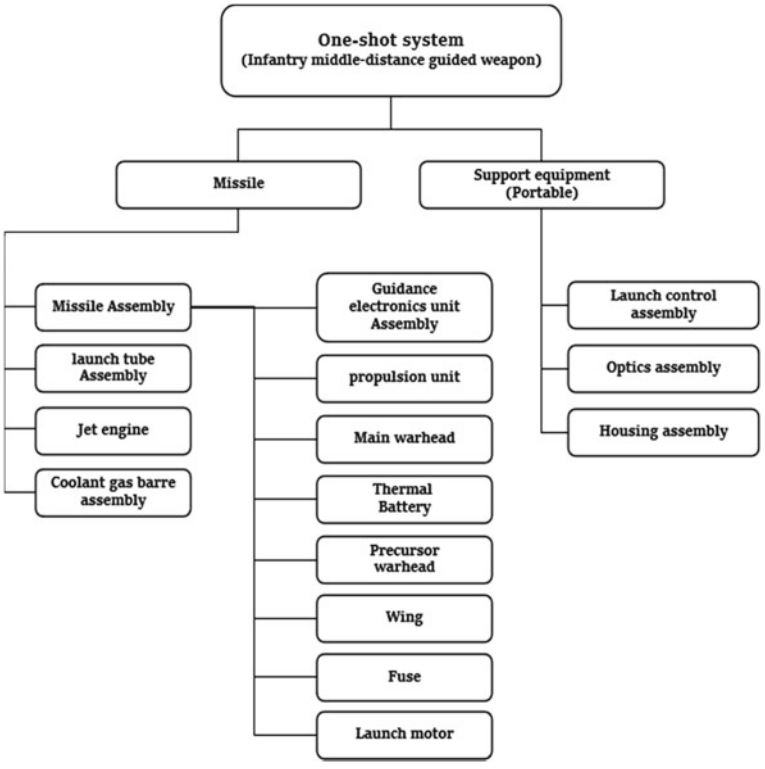


Fig. 1 An example of one shot system with regular equipment

type 1 unit follows a Weibull distribution with an increasing failure rate, the failure possibility of units may increase over time. To improve the system performance (system availability), it is useful to maintain the aging working units preventively at the inspection points, replacing the old type 1 unit that reached the preventive maintenance threshold with a new one. Type 2 units in one shot systems can be ejection unit, propulsion unit, main warhead, precursor warhead, thermal battery, and fuse, and their state deteriorates gradually and degradation beyond specific limits makes the system incapable of normal operation. Generally, the failures of type 2 units can only be detected by destructive testing and the predetermined useful lifetime which is determined based on their reliability test in the development phase is given in advance. For example, the lifetime of the thermal battery is 10 years, and the other type 2 units in a missile are about 14–15 years. Type 2 units experience degradation during their storage period due to corrosion, thermal fatigue, or repeated shock load. It is hard to know the exact state of type 2 units, the degradation function is assumed to be known and the units fail when the degradation level exceeds a failure threshold. For preventing failures, the type 2 units whose degradation level reaches a preventive maintenance threshold are replaced by new ones. Brownian motion, compound poisson process [24], gamma process [25, 26], and randomized degradation functions proposed by Li and Pham [27] are widely used for modeling degradation in the reliability area. Some one-shot systems consist of one-shot devices and regular equipment together and the one shot system functions normally on demand when the one shot device and regular equipment work together [22]. For example, a missile system needs support equipment to launch a missile and it is also composed of several units in series structure. As shown in Fig. 1, support equipment units such as the launch control, optics, and housing assembly in the launchers fail randomly, and the failure of units can be repaired upon failure because the failure can be detected immediately. The failure times of units follow some specific failure distributions, just like the type 1 units in one shot devices. Without the assistance of support equipment, a one shot device cannot operate even if it is in a working state on demand.

3 Storage Reliability (Availability) of One Shot Systems

The systems like missiles, airbags, and fire extinguishers are required to survive prolonged periods of storage and perform a function successfully once only if needed. The ability of systems in a dormant state to keep their required functions is measured by storage reliability [28]. The reliability of a one shot system is defined as the probability that it performs the required function only once, and only when demanded, under stated conditions and for a specified period of time [19]. System reliability has been widely used as a system performance measure to assess storage systems [23]. The assessment of system storage reliability has been widely studied for several decades; see Merren [29], Zhao and Xie [30], Zhao et al. [31]; Su et al. [32], Zhang et al. [33], Liu and Liu [34]; Zheng and Xin [22].

In particular, Zhang et al. [28] classified the analysis approach of storage reliability into three categories, namely physics-of-failure approaches, analytical modeling involving lifetime data, and accelerated storage test measures. Jang and Son [22] proposed a reliability estimation model of one shot system using quantal-response data expressed as success or failure numbers in samples to be tested at an arbitrary time point in the destructive inspection, and these data usually are collected by testing of one shot systems such as missile and airbag. Parametric, nonparametric, and Bayesian methods can be used to estimate reliability based on the quantal-response data. In destructive testing, it is difficult to collect sufficient lifetime information on the devices, an accelerated life-test (ALT) is usually performed in order to collect more failures in a short period of time [35–37], and usually, the lifetime distribution of the system is also unknown [38, 39]. Thus, the estimation problem of the system reliability of one shot system is the key but difficult one in the reliability research area. The reliability of the one shot system is affected by aging or degradation during storage periods and its reliability declines over time. To meet its high-reliability requirements, the system is periodically inspected, and the failed units are maintained. For these repairable systems, system availability is a more appropriate index to evaluate system performance.

Inspection models for systems in which system failures can be detected only by inspection have been studied by many researchers in the reliability area [40]. Recently, Sarkar and Sarkar [41] studied the instantaneous availability and steady-state availability of a safety system, which is maintained under periodic inspection and a perfect repair policy with constant repair time. Cui and Xie [42] did similar work by considering random downtime caused by repair. Tang et al. [43] investigated the system availability of a protective system considering non-negligible downtime and they obtained instantaneous availability and steady-state availability of a periodically inspected system with multiple failure modes. Qiu et al. [44] derived instantaneous availability and steady-state availability for a system suffering from hidden failures and steady-state availability is considered an optimization criterion to obtain the optimal inspection interval.

Most works on instantaneous and steady-state availability of protective and safety equipment assumed hidden failures and they studied optimal inspection policies to detect failures. However, On the other hand, Martinez [45] first proposed a storage reliability problem for predicting the reliability of devices that must be in storage or dormant period for long time periods prior to usage. Storage reliability is defined as "the probability a device will perform its intended function after several years of storage," and these devices are periodically tested while in storage. Even though the storage reliability model of Martinez [45] described well the characteristics and situations of many real one shot systems, few studies refer to this model when dealing with storage reliability (availability) and inspection problems. Kim and Yun [2] derived the instantaneous availability models before and after inspection respectively, when the time to system failure follows exponential and Weibull distributions. Perfect repair is considered in the exponential case, and minimal repair is considered in the Weibull distribution case. They also assumed that undetected failures cannot be detected until the system reaches its lifetime.

In general, the steady-state (average/mean/limiting) availability is defined as the average proportion over a long time period when the system can operate, and it is usually used as a main reliability measure to evaluate the performance of complex repairable systems. Kitagawa et al. [3] assumed minimal repair is performed when a failure is detected, and the one shot system is replaced when the nth failure is detected. They derived the mean availability for a single-unit one shot system. The duration between two successive replacements is regarded as one cycle and the downtime due to inspection, minimal repair, and replacement is considered. Kitagawa et al. [4] formulated the mean availability for a sing-unit one shot system when it is replaced after *n*th failure and at the *m*th periodic inspection point, respectively. However, since one shot system performs tasks suddenly at an unknown time point, (for example, the pop-up airbags in a car accident), availability during finite horizon is more suitable as an optimization criterion to analyze and optimize the one shot systems. Hence, interval availability which is the average probability that the one shot system is in a functioning state between the inspection times can be used as a measure to evaluate the performance of one shot systems. Yun et al. [21, 26], and Zhao and Yun [5] considered interval availability as the average proportion of time where the one shot system is functioning within a given inspection interval.

Table 1 summarizes the various availability models for one shot systems that are studied in existing reliability and availability models. Based on the operation model, methodology, and performance measure, the existing models can be categorized. Three papers considered four phases in the operation situation of one shot systems and some papers estimated the system availability by simulation. Instantaneous and interval availabilities were used as the system performance measurement. Monte Carlo and discrete event simulation models are used to evaluate the interval availability of one shot systems with complex structures.

In this section, we introduce basic storage reliability (instantaneous availability) models for a single-unit one shot system with several operation phases and inspection errors in Sect. 3.1. In Sect. 3.2 an interval availability model based on periodic inspection for a multi-unit one shot system is also introduced.

3.1 Instantaneous Availability for a Single- Unit One Shot System

In this sub-section, we introduce two basic instantaneous availability models in which the one shot system has four operation phases. The one shot system is inspected (tested) periodically to detect the failures of parts and the parts that failed are sent back to storage after repair. Figure 2 shows the four phases and expresses an operation model that determines the storage reliability of the system after the system is produced in the factory and before use on demand. The system operation procedure in Martinez [45] is divided into factory test (Phase 1), Shipping transportation handling (Phase 2), Pre-acceptance checkout (Phase 3), and Storage (periodic inspection, Phase 4)

Table 1 Summary of contributions to one shot system availability

Contribution	System structure	Operation model	Methodology	Performance measure
Martinez [45]	Single system; Single unit	Four phases	Mathematical	Instantaneous availability (reliability)
Cui et al. [1]	Single system; Single unit	One phase	Mathematical	Instantaneous availability and Limiting average availability
Yun et al. [21]	Single system; Multi-unit; Series	One phase	Mathematical and discrete-event simulation	Interval availability
Yun et al. [26]	Single system; Multi-unit; Series	One phase	Discrete-event simulation	Interval availability
Kim and Yun [2]	Single system; Single unit	Four phases	Mathematical	Instantaneous availability
Kitagawa et al. [3]	Single system; Single unit	Four phases	Mathematical (descent method)	Mean availability
Kitagawa et al. [4]	Single system; Single unit	One phase	Mathematical (approximation)	Mean availability
Liu et al. [17]	Multiple system; Multi-unit; Series	One phase	Discrete-event simulation	Interval availability
Zhao and Yun [18]	Multiple systems; Multi-unit; Series	One phase	Discrete-event simulation	Interval availability
Zhao and Yun [5]	Single system; Multi-unit; Series	One phase	Mathematical and Monte Carlo simulation	Interval availability

before use, and in phase of storage, periodic inspection is performed. Each phase has a different time period and failure rate and the failed system detected by inspection is maintained and returned to storage.

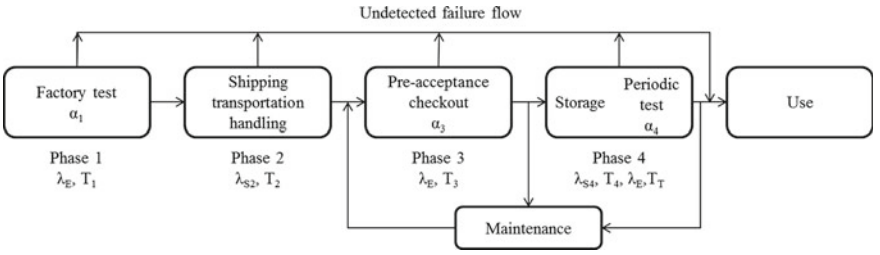


Fig. 2 Storage reliability model under periodic inspection [45]

3.1.1 Martinez Storage Availability Model

Martinez [45] calculated the number of expected failures in the pre-storage stage as the sum of expected failures occurring in Phases 1–3 as shown in (Eq. 1), considering inspection errors.

$$F_3 = (1 - \alpha_1)\lambda_E T_1 + (1 - \alpha_3)(\lambda_{S2} T_2 + \lambda_E T_3) \tag{1}$$

Then, the system reliability before storage is

$$R = e^{-F_3} \tag{2}$$

During the T_4 period, the one shot system stays in storage, and then the system is powered on for inspection during the T_T period. After inspection, the system is turned off and goes back to storage. Then, the time to turn the system on and off for inspection is the time the system is operated during the storage period. Hence, the expected number of failures during one cycle in the storage state is given in (Eq. 3).

$$F_P = \lambda_{S4} T_4 + \lambda_E T_T + \lambda_C C \tag{3}$$

where λ_C is the power on–off cycling failure rate and C is the number of test cycles in the test interval. Using (Eqs. 1 and 3), the system reliability is calculated before and after Nth inspection.

$$R_{N(\text{Min})} = e^{-[(N-1)(1-\alpha_4)F_P + F_3]} \times e^{-\lambda_{S4} T_4} \tag{4}$$

$$R_{N(\text{Max})} = e^{-[N(1-\alpha_4)F_P + F_3]} \tag{5}$$

3.1.2 Kim-Yun Instantaneous Availability Model

Kim and Yun [2] modified the storage reliability model under periodic inspection in the one shot system of Martinez [45] in which the failures of one shot system during four operation phases follow exponential distributions, which can be applied to a system with a general failure distribution model. System instantaneous availability can be obtained by calculating the reliability of each phase. Let R_i denote the probability that the system is working after the inspection of phase i. Thus, R_1 is the probability that the system is passed at factory test, which includes the probability that the system is not failed until inspection at the factory, and the probability that the system fails, but the failure is detected and maintained to a new one. R_2 is the probability that the system survives at the pre-acceptance checkout, including the probability that the system survives until arrival at the storage location after transportation and the probability that the system is failed, the failure is detected and

maintained to a new one. In addition, R_3 is the probability that the system survives at periodic tests after long-term storage, which includes three types of probabilities. The first one is the probability that the system survives the test after storage, the second one is the probability that the failure does not occur at the C times on–off tests of inspection equipment, and the last one is the probability that the failures are detected by inspection and maintained to a new one. Exponential and Weibull distributions are considered as the distribution of the time to system failure.

3.1.3 Exponential Case

If the failure time of the system follows an exponential distribution, reliability before and after the test can be obtained.

$$R_1 = e^{-\lambda_E T_1} + \alpha_1 (1 - e^{-\lambda_E T_1})$$

$$R_2 = e^{-\lambda_{S_2} T_2 - \lambda_E T_3} + \alpha_3 (1 - e^{-\lambda_{S_2} T_2 - \lambda_E T_3})$$

$$\begin{aligned} R_3 = {} & e^{-\lambda_{S_4} T_4 - \lambda_E T_T} \cdot (1 - \lambda_c)^C + \alpha_4 \\ & \Big\{ (1 - e^{-\lambda_{S4} T_4}) + e^{-\lambda_{S4} T_4} \cdot (1 - e^{-\lambda_E T_T}) \\ & + e^{-\lambda_{S4} T_4 - \lambda_E T_T} \Big(\sum_{i=1}^{C-1} \lambda_C (1 - \lambda_C)^k \Big) \Big\} \end{aligned} \tag{6}$$

Storage availability (R_N) is defined as the probability that the system is working at Nth inspection after storage. Thus, the storage availabilities before and after Nth inspection, $R_{N(\min)}$ and $R_{N(\max)}$, are derived.

$$\begin{aligned} R_{N(\min)} &= (R_1 \cdot R_2 \cdot R_3^{N-1}) \cdot e^{-\lambda_{S_4} t} \\ R_{N(\max)} &= R_1 \cdot R_2 \cdot R_3^N \end{aligned} \tag{7}$$

3.1.4 Weibull Distribution Case

When the time to failure of the one shot system at each phase has a Weibull distribution with the same shape parameter. In order to derive the system availability, the cumulative exposure model (CEM) to failure distributions at different phases is assumed [46]. When the system is failed, the failure is maintained at the inspection, and the system is assumed to be repaired minimally (minimal repair assumption) and be restored to the same condition as the condition before the failure.

The probabilities that the system is working at the inspection of phase $i (i = 1, 2, 3)$ are given

$$R_1 = e^{-(\lambda_E T_a)^\beta} + \alpha_1\left\{1 - e^{-(\lambda_E T_a)^\beta}\right\}$$

$$R_2 = \frac{e^{-\left\{\lambda_{s_2}\left(T_a^*+T_b\right)\right\}^\beta - \left\{\lambda_E\left(T_b^*+T_c\right)\right\}^\beta}}{e^{-(\lambda_{s2}T_a^*)^\beta - (\lambda_E T_b^*)^\beta}} + \alpha_3\left(1 - \frac{e^{-\left\{\lambda_{s_2}\left(T_a^*+T_b\right)\right\}^\beta - \left\{\lambda_E\left(T_b^*+T_c\right)\right\}^\beta}}{e^{-(\lambda_{s2}T_a^*)^\beta - (\lambda_E T_b^*)^\beta}}\right)$$

$$\begin{aligned} R_3 = {} & \frac{e^{-\left\{\lambda_{s_4}\left(T_c^*+T_s\right)\right\}^\beta - \left\{\lambda_E\left(T_s^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_c^*)^\beta - (\lambda_E T_s^*)^\beta}} \times (1-\lambda_C)^C \\ & + \alpha_4\left[\left(1 - \frac{e^{-\left\{\lambda_{s_4}\left(T_c^*+T_s\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_c^*)^\beta}}\right) + \frac{e^{-\left\{\lambda_{s_4}\left(T_c^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_c^*)^\beta}} \times \left(1 - \frac{e^{-\left\{\lambda_E\left(T_s^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_E T_S^*)^\beta}}\right)\right. \\ & \left. + \frac{e^{-\left\{\lambda_{s_4}\left(T_c^*+T_T\right)\right\}^\beta - \left\{\lambda_E\left(T_d^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_c^*)^\beta - (\lambda_E T_s^*)^\beta}} \times \sum_{k=0}^{C-1} \lambda_C(1-\lambda_C)^k\right] \end{aligned} \tag{8}$$

where $T_a^* = \frac{\lambda_E}{\lambda_{s2}}T_a$, $T_b^* = \frac{\lambda_{s2}}{\lambda_E}\left(T_a^* + T_b\right)$ $T_c^* = \frac{\lambda_E}{\lambda_4}\left(T_b^* + T_c\right)$, $T_s^* = \frac{\lambda_{s4}}{\lambda_E}\left(T_c^* + T_s\right)$

Thus, the availability just before and after Nth inspection are:

$$R_{N(Min)} = \begin{cases} \begin{aligned} & R_2 \times \frac{e^{-\left\{\lambda_{s_4}\left(T_c^*+T_s\right)\right\}^\beta - \left\{\lambda_E\left(T_s^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_c^*)^\beta - (\lambda_E T_s^*)^\beta}} + \left(1 - \frac{e^{-\left\{\lambda_{s4}\left(T_c^*+T_s\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_c^*)^\beta}}\right) \\ & \quad + \left(\frac{e^{-\left\{\lambda_{s4}\left(T_c^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_E T_s^*)^\beta}} \times \left(1 - \frac{e^{-\left\{\lambda_E\left(T_s^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_E T_s^*)^\beta}}\right)\right), \quad N = 1 \\ & R_{N-1} \times \frac{e^{-\left\{\lambda_{s_4}\left(T_{T_{N-1}}^*+T_s\right)\right\}^\beta - \left\{\lambda_E\left(T_{s_{N-1}}^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_{T_{N-1}}^*)^\beta - (\lambda_E T_{s_{N-1}}^*)^\beta}} + \left(1 - \frac{e^{-\left\{\lambda_{s4}\left(T_{T_{N-1}}^*+T_s\right)\right\}^\beta}}{e^{-(\lambda_{s4}T_{T_{N-1}}^*)^\beta}}\right) \\ & \quad + \left(\frac{e^{-\left\{\lambda_{s4}\left(T_{s_{N-1}}^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_E T_{s_{N-1}}^*)^\beta}} \times \left(1 - \frac{e^{-\left\{\lambda_E\left(T_{s_{N-1}}^*+T_T\right)\right\}^\beta}}{e^{-(\lambda_E T_{s_{N-1}}^*)^\beta}}\right)\right), \quad N \geq 2 \end{aligned} \end{cases} \tag{9}$$

$$R_{N(\text{Max})} = \begin{cases} R_3, & N = 1 \\ R_{sN}, & N \geq 2 \end{cases} \tag{10}$$

3.2 *Interval Availability for a Multi-Unit One Shot System*

In the previous section, we introduced the instantaneous availability models of a single-unit one-shot system with multi-phases. In this section, we consider a multi-unit one-shot system with a series structure. Zhao and Yun [5] proposed a storage availability (interval availability) model for a multi-unit one shot system with an imperfect inspection. The failure of each unit in the system can be detected with a

Fig. 3 Interval availability of one shot system [5]

constant probability, α. In their model, the one shot system is perfect and all units in the one shot system are working at the starting point of the first storage. The one shot system is inspected periodically and there are two cases in which undiscovered failures can be detected at the following inspections or are no longer detected. For the one shot system under periodic inspection, the interval availability is proposed to evaluate the performance of the one shot system. The interval availability, $A(k, T)$, is defined as the mean proportion of uptime within the inspection interval under the condition that the one shot system is working at the starting time point, as shown in Fig. 3.

The interval availability between kT and $(k + 1)T$ is

$$\begin{aligned} A(k,\ T) &= \frac{UT(k,\ T)}{T} = \frac{y_1(k,\ T)\int_{kT}^{(k+1)T} R(x,\ T)dx}{T} \\ &= y_1(k,\ T) \cdot A_{av}(T) \end{aligned} \tag{11}$$

where $R(x, T)$ is the reliability at x when the inspection interval is T, the $A_{av}(T)$ is the expected fraction of operating time during a given interval T. The $y_1(k, T)$ represent the probability that the system working at the time kT, $k = 1,2, \ldots, K$, which is calculated differently, depending on the assumption whether the discovered failure can be detected again.

Thus, the interval availability between 0 and kT, can be calculated from $A(i, T)$, $i = 0, 1, \ldots, k - 1$. When the failure times of units follow exponential distributions with failure rates, λ_i, $i = 1,..,n$, $A_{av}(T)$ is

$$A_{av}(T) = \frac{\frac{1}{\sum_{j=1}^{n}\lambda_j}\left(1 - e^{-\sum_{j=1}^{n}\lambda_j T}\right)}{T} \tag{12}$$

Because the one-shot system has a series structure.

3.2.1 Case in Which Undiscovered Failures Cannot Be Detected Eventually

The system is perfect before the first storage, and the probability is $y_1(0, T) = 1$. When $k > 1$, $y_1(k, T)$ is the achievement of two independent probabilities, namely the probability of the system working at the time $(k\text{-}1)T$, $y_1((k - 1), T)$, and the

probability of the system being functional after interval T, $p_{(1,1,\ldots,1)}$:

$$y_1(k,T)=\begin{cases}1, & k=0\\ y_1((k-1),\ T)\times p_{(1,\ 1,\ldots,1)}, & k>0\end{cases} \tag{13}$$

The state vector (1, 1, …, 1) means that all units are in a working state, and the probability of state vector (1, 1, …, 1) after one cycle T is

$$\begin{aligned}p_{(1,1,\ldots,1)}=&\prod_{j=1}^{n}R_j(T)+\alpha F_1(T)\prod_{j=2}^{n}R_j(T)+\cdots+\alpha^{n-1}\prod_{j=1}^{n-1}F_j(T)R_n(T)\\&+\alpha^n\prod_{j=1}^{n}F_j(T)\end{aligned} \tag{14}$$

where $R_j(T)$ is the storage reliability of unit j at time t. and α is the probability that a failures is detected.

3.2.2 Case in Which Undiscovered Failures Can Be Detected Again

This case considers the probability that the failures being undiscovered at the $(k-1)$th inspection, detected and repaired at the kth inspection, which is denoted by $(z(k-1), T)$.

$$y_1(k,T)=\begin{cases}1, & k=0\\ \prod_{j=1}^{n}R_j(T)+\alpha F_1(T)\prod_{j=2}^{n}R_j(T)+\ldots & \\ +\alpha^{n-1}R_1(T)\prod_{j\,=\,2}^{n}F_j(T)+\alpha^n\prod_{j=1}^{n}F_j(T), & k=1\\ y_1((k-1),\ T)\cdot p_{(1,\ 1,\ldots,1)}+z((k-1)T), & k>1\end{cases} \tag{15}$$

where

$$\begin{aligned}z((k-1),\ T)=&\,y_{0,1,\ldots,1}((k-1),T)\alpha\prod_{j=2}^{n}R_j(T)+\cdots+y_{0,0,\ldots,1}((k-1),T)\alpha^{n-1}R_n(T)\\&+\Bigg(y_{0,1,\ldots,1}((k-1),T)\prod_{j=2}^{n}F_j(T)+\cdots+y_{0,0,\ldots,1}((k-1),T)F_n(T)\\&+y_{0,0,\ldots,0}((k-1),T)\alpha^n\Bigg)\end{aligned}$$

Since there are n units in the model, and the unit state x_i ($i=1,2,..,n$) also has '0' and '1' state, then $y_{0,\,1,\ldots,1}((k\text{-}1),\ T)$ is the probability that the system fails and the units are in the state (0, 1, …, 1) at $(k-1)T$.

For problems where it is difficult to calculate interval availability analytically, researchers also used the Monte Carlo [5] or discrete event simulations [17, 21, 26] to evaluate the interval availability when one shot systems having multiple units, complex failure modes and maintenance processes are under consideration.

4 Inspection Policies of a Single One Shot System

In Sect. 3, we introduced the instantaneous availabilities of a single-unit one shot system and the interval availability of a multi-unit one-shot system. One shot system like airbags needs to ensure that their functions are successfully implemented as needed throughout the lifetime, regardless of their storage time. However, during storage, the electric and chemical parts of the system can be affected by many kinds of environmental factors such as temperature, humidity, and mechanical stresses. The one shot system can fail to perform the function required on demand because some kinds of electronic and electric parts degrade over time [47, 48]. The state of the one shot system only can be observed when the system is operating or inspected. Therefore, maintenance-related actions (inspection, repair, and replacement) are required to ensure high storage reliability throughout the lifetime of the one shot system. Conducting either periodical or continuous inspection is an effective way of observing the state of standby [49, 50], safety [51], protection [51], hidden failure [43, 44] or storage system and assessing their degradation. Nakagawa [40] reviewed and summarized several inspection models 413 for systems with standby units. Nakagawa and Mizutani [52] reviewed three inspection models over a finite time span: periodic inspection, sequential inspection, and asymptotic inspection.

For the one shot system, Ito and Nakagawa [7] considered the periodic inspection policy for a one shot system with two units. The system is inspected periodically until the system reliability is equal to or lower than a predetermined value q, at which time the system is overhauled. After each inspection, unit 1 is maintained as a new one, while the failure rate of unit 2 remains unchanged. An optimal inspection time that minimizes average cost, including inspection and overhaul costs, is obtained. Ito et al. [8] further extended the work of Ito and Nakagawa [7] and considered unit 2 that consists of units 2–1 and 2–2. Unit 2–1 is replaced at the inspection, but unit 2–2 is not replaced. Expected total cost, including inspection cost and downtime (time elapsed between failure and its detection) cost, is considered as an optimization criterion to determine the optimal inspection interval. Later, Ito and Nakagawa [11] assumed that the system is replaced at inspection time points or before failure when its reliability becomes lower than q. In the periodic inspection policy of Ito and Nakagawa [12], unit 2 is assumed to have two failure rate functions, where one is degraded with time, and the other one is degraded at each inspection. Ito and Nakagawa [6] further extended this model by considering the expected cost rate model. Then they obtained the optimal inspection time that maximizes overhaul time and minimizes the expected cost rate. Cui et al. [1] derived instantaneous availability and limiting average availability models under periodic inspection for a single-unit

storage system. Kitagawa et al. [9] proposed a periodic inspection policy for a multi-unit one shot system. When a failure is detected, the unit is minimally repaired, and all units are replaced when nth failure is detected. Kitagawa et al. [4] proposed and compared number-of-failure-based (NOFB) and periodic policies for a single-unit one shot system with the periodic inspection.

However, during the inspection, the age of the unit changes after each maintenance action, which depends on the maintenance model of each unit. This information should be used to decide upon the timing of the next inspection. Therefore, aperiodic inspection is more appropriate for the general distribution model. Nakagawa et al. [13] assumed that the system deteriorates with age and fails according to general distribution. They summarized two inspection policies (periodic and sequential) by considering a random working time. Then, they proposed an inspection policy that the system is inspected at every completion of the nth working times. Finally, a back review model that the system goes back to the latest inspection time when it has failed is appeared. Yun et al. [21] formulated an interval availability model to obtain an optimal inspection interval under the periodic policy. In addition, for a general distribution model, they used simulation to estimate the interval availability and life cycle cost under the aperiodic policy. Kitagawa et al. [3] also formulated an expected cost rate and mean availability under periodic and aperiodic inspection policies.

Considering the effect of inspection, the inspection can be divided into perfect and imperfect ones. In most existing inspection models, it has been assumed that the inspection is perfect and can identify all system failures. However, in practice, Imperfect inspection is a common situation observed in the industry. Due to some operation errors, the aging of inspection equipment and other uncertain factors, the inspection may give a wrong result. Some failures in imperfect inspection remains unrevealed, and a functioning system is regarded as a failure; some researchers considered this practical situation in their inspection models; see Kaio and Osaki [53], Badia et al. [54], and Berrade et al. [55]. For one shot system, few studies consider inspection errors in their inspection models. In the availability models of Martinez [45] and Kim and Yun [2], due to inspection error, some failures remain unrevealed during all phases. Cho and Lee [10] presented a two-stage inspection policy in which the system is inspected by a simple inspection at periodic times. Later, a precise inspection was carried out to confirm the final state of the system if the result of the simple inspection showed that the system is in a state of failure. The probability of two types of errors were considered in the simple inspection: one is that an operating system might be regarded as failing with a constant probability a ($0 \leq a \leq 1$) and the other is the failure remains undiscovered with a constant probability b ($0 \leq b \leq 1$).

This section summarizes, classifies, and compares inspection policies of one shot systems. Table 2 lists the various inspection policies for a single one shot system that is summarized from a number of existing optimal inspection models. From Table 2, Yun et al. [21] studied an inspection policy problem for a one shot system with two types of units over a finite time span, in which type 1 units should be maintained at inspection and type 2 units should be replaced by new ones at predetermined times.

They proposed a heuristic method to find an optimal inspection interval, which satisfies a target interval availability when the time to failure of type 1 units follows an exponential distribution. In addition, when the failure time of type 1 units follows different distributions, a simulation-based optimization procedure with a genetic algorithm was proposed to find global optimal inspection times for the one shot system. Later, Yun et al. [26] considered a gamma process to model the degradation process of type 2 units and apply a gamma-bridge sampling method for simulating the gamma process. Preventive maintenance policies for type 1 and 2 units are considered to improve system interval availability. Zhao and Yun [5] recently proposed an imperfect model based on periodic inspection, and they assumed that the probability for failure identification, α, is a constant value. They derived interval availability and life cycle cost models that are considered as optimization criteria to obtain the optimal inspection interval analytically. They also used Monte-Carlo simulation to evaluate the interval availability and life cycle cost when a large-sized system is under consideration. A heuristic algorithm is proposed to find the optimal solutions.

This section explores the optimation problems to determine the optimal inspection interval or inspection time points of the one shot systems studied so far and reviews the typical methodologies. Section 4.1 introduces inspection policies for the single-unit one shot system. Section 4.2 introduces inspection policies for the multi-unit one shot systems.

4.1 Optimal Inspection Policies for Single-Unit One Shot Systems

In this subsection, we introduce periodic and sequential inspection policies of single-unit one shot systems in Kitagawa et al. [3]. In the paper, a single-unit one shot system is inspected with equal interval (periodic inspection model) or different intervals (sequential inspection model). The failure distribution of the one shot system is IFR (increasing failure rate). When the failure of the system is detected, the system is repaired minimally. At the nth failure of the system, the system is replaced by a new one. The objective is to find the optimal number of failures before replacement and the inspection intervals to minimize the expected cost rate ensuring a specified mean availability.

4.1.1 Periodic Inspection Policy with Minimal Repair

An optimal inspection interval and the number of minimal repairs before replacement are determined, which minimizes the expected cost rate and ensures a specified mean availability. The mean availability and expected cost rate under periodic inspections are given.

Table 2 Summary of contributions on optimal inspection models for a single one shot system

Contribution	System structure	Inspection policy	Optimization criteria	Methodology	Decision variables
Ito and Nakagawa [7]	Single system; Two units	Periodic	Expected cost rate	Mathematical	Optimal inspection time; Time to overhaul
Ito and Nakagawa [11]	Single system; Two units	Periodic	Total expected cost until a replacement	Mathematical	Optimal inspection time
Ito and Nakagawa [12]	Single system; Two units	Periodic	Expected total cost until the detection of failure	Mathematical	Optimal inspection time
Ito et al. [8]	Single system; Three units	Periodic	Expected cost rate	Mathematical	Optimal inspection time; Optimal inspection number; Optimal replacement number
Ito and Nakagawa [6]	Single system; Two units	Periodic	Overhaul time Expected cost rate	Mathematical	Optimal inspection times T^*
Cho and Lee [10]	Single system; Single-unit	Imperfect Periodic	Expected cost rate	Mathematical	Optimal inspection time T^* and number of inspections N^*
Cui et al. [1]	Single system; ingle-unit	Periodic	Expected cost until the detection of failure	Mathematical	Instantaneous availability and Limiting average availability
Yun et al. [21]	Single system; Multi-unit; Series	Periodic Aperiodic; Replacement	Interval availability; Life cycle cost	Mathematical and Simulation model; GA and Heuristic	Optimal inspection intervals

(continued)

Table 2 (continued)

Contribution	System structure	Inspection policy	Optimization criteria	Methodology	Decision variables
Yun et al. [26]	Single system; Multi-unit; Series	Aperiodic	Interval availability; Life cycle cost	Simulation model; GA and Heuristic	Inspection intervals; Preventive replacement ages; Preventive maintenance thresholds
Kitagawa et al. [9]	Single system; Multi-unit;	Periodic	Expected cost rate	Mathematical; Monte-Carlo simulation	Optimal inspection interval
Kitagawa et al. [3]	Single system; Single-unit	Periodic and aperiodic Minimal repair	Mean availability; Expected cost rate	Mathematical; Simulated annealing	Optimal number of failures until replacement; Optimal inspection interval
Kitagawa et al. [4]	Single system; Single-unit	Periodic	Mean availability; Expected cost rate	Mathematical	Number of failures until replacement or number of inspections; Optimal inspection interval;
Zhao and Yun [5]	Single system; Multi-unit; Series	Imperfect; Periodic	Interval availability; Expected life cycle cost	Mathematical and Monte-Carlo simulation	Optimal inspection interval; Number of inspections

$$A_{av}(n,T) = \frac{\sum_{l=1}^{n} \mu^{(l)}}{(T+m_I)\sum_{l=1}^{n}\sum_{k=0}^{\infty}(kT)+(n-1)m_R+m_P} \tag{16}$$

$$C(n,T) = \frac{C_I \sum_{l=1}^{n}\sum_{k=0}^{\infty}\overline{F}^{(l)}(kT)+(n-1)C_R+C_P}{(T+m_I)\sum_{l=1}^{n}\sum_{k=0}^{\infty}\overline{F}^{(l)}(kT)+(n-1)m_R+m_P} \tag{17}$$

where C_I, C_R and C_P refer to the inspection, minimal repair, and replacement costs, respectively; and m_I, m_R and m_P refer to the mean duration of inspection, minimal repair, and replacement, respectively. The $\mu^{(l)}$ denote the mean operating time between*(l-1)*th failure and *l*th failure.

For a periodic inspection policy, the mean availability is unimodal with respect to inspection interval in many cases when the failure occurrences follow Weibull

distribution. Let T^* be the inspection interval that gives the maximum mean availability. Then, the expected cost rate is strictly decreasing for $T > T^*$. Let T_1 and T_2 be inspection intervals that satisfy $A_{av}(n,T_1) = A_{av}(n,T_2) = A_T$ and $T_1 < T_2$. A heuristic method is proposed to find the optimal T^* for given n as follows.

Step 1 Obtain T^* using the golden section method.
Step 2 If $A_{av}(n, T^*) < A_T$, the optimal solution does not exist. Otherwise, get T_1, T_2 and $C(n, T_2)$.
Step 3 Get a minimal expected cost rate for $[T_1, T^*]$ using a descent method, and compare it to $C(n, T_2)$. The lower one is the minimal expected cost rate and T is optimal.

Then, we change the value of n, compare the expected cost rates and finally determine the optimal (n, T).

4.1.2 Sequential Inspection Policies with Minimal Repair

In the sequential (non-periodic) inspection model, the inspection interval is changed after the failure of the system is detected and the system is repaired minimally. $T^{(l)},\ldots,T^{(n)}$ are inspection intervals between minimal repairs and the system is replaced when nth failure is detected. If the failure time of the system follows a Weibull distribution with scale parameter η and shaper parameter β ($\beta > 1$) and the system failures can be detected immediately, the mean operation time between the (l-1)th failure and the lth failure is

$$\mu^{(l)} = \int_0^\infty \frac{[H(t)]^{l-1}}{(l-1)!} e^{-H(t)} dt = \frac{\eta}{\beta} \frac{\Gamma(l+1/\beta-1)}{\Gamma(l)} \tag{18}$$

where $H(t)$ is the cumulative hazard rate function. The mean availability of the system and expected total cost rate are given.

$$A_{av}(T^{(1)},\ldots,T^{(n)}) = \frac{\sum_{l=1}^{n} \mu^{(l)}}{\sum_{l=1}^{n} \sum_{k=0}^{\infty} \left\{ (T^{(l)} + m_I) \overline{F}^{(l)}(kT^{(l)}) \right\} + (n-1)m_R + m_P} \tag{19}$$

$$C(T^{(1)},\ldots,T^{(n)}) = \frac{C_I \sum_{l=1}^{n} \sum_{k=0}^{\infty} \overline{F}(l)\left(kT^{(l)}\right) + (n-1)C_R + C_P}{\sum_{l=1}^{n} \sum_{k=0}^{\infty} \left\{ (T^{(l)} + m_I) \overline{F}^{(l)}(kT^{(l)}) \right\} + (n-1)m_R + m_P} \tag{20}$$

In order to obtain the optimal n and $T^{(l)},\ldots,T^{(n)}$, a simulated annealing algorithm is used and the procedure for a fixed n is shown as follows.

Step 1 Choose an initial solution $\{T^{(1)}, \ldots, T^{(n)}\}$ that satisfies $A_{av} \geq A_T$, and calculate its expected cost rate.

Step 2 Change inspection intervals randomly from each present interval within + 3.0.

Step 3 Calculate the expected cost rate and mean availability. If the new availability satisfies $A_{av} \geq A_T$ and the new expected cost rate is lower than the old one, renew the solution. However, if the new availability satisfies $A_{av} \geq A_T$ and the new expected cost rate is larger, renew the solution with probability $\exp(-D/t)$, where t is temperature and D is the difference in expected cost rate. Otherwise reject the new solution.

Step 4 If the solution is not improved during 200 times repetitions in a series, determine the optimal solution as the present solution. Otherwise, multiply t by 0.95 (geometric cooling schedule) and go to step 2.

Then, change the value of n, compare the expected cost rates and finally determine the optimal $(n, T^{(1)}, \ldots, T^{(n)})$.

4.2 Optimal Inspection Policies for Multiple Units One Shot Systems

In this subsection, we introduce periodic and sequential inspection policies for multi-unit one shot systems. In the paper, multi-unit one shot systems are inspected with equal intervals (periodic inspection model) or different intervals (sequential inspection model). Some units can be maintained and other units cannot be maintained in multi-unit one shot systems at inspection. Additionally, the system is replaced based on pre-specified conditions, for example, at periodic times NT to hold high reliability. The objective is to find the inspection intervals to minimize the expected cost rate ensuring a specified system performance measure (for example, availability).

4.2.1 Periodic Inspection Policy of a One Shot System with Two Types of Units

Ito and Nakagawa [7] proposed a basic inspection optimization model for a multi-unit one shot system They assume that a one shot system consists of two types of units and unit i has a cumulative hazard function $H_i(t)(i = 1, 2)$. The system is inspected and maintained at periodic times NT $(N = 1, 2\ldots.)$ and is overhauled if the reliability is small or equal to q. Unit 1 is maintained and becomes new after every NT and unit 2 is not done, i.e., its hazard rate remains unchanged by any inspection.

Inspection and overhaul costs are considered, and the expected cost rate is derived. The optimal inspection interval T and inspection number N to minimize the expected cost rate are obtained from the required system reliability, q. The time to overhaul is $NT + t_0$, where $t_0(0 < t < T)$ and $exp[-H_1(t_0) - H_1(t_0)] = q$ should be satisfied.

The expected cost rate $C(T) = Nc_1 + c_1/NT + t_0$, and the optimal inspection interval T minimizes the expected cost rate and satisfy the equation. Ito and Nakagawa [11, 12] proposed a periodic inspection of a one shot system with two types of units in which unit 1 is inspected and maintained at each inspection, but unit 2 is not done. The system is replaced at the inspection of failure or at the time when the reliability is below pre-specified reliability. Ito et al. [8] considered three types of units in a one shot system in which unit 1 is maintained at each inspection, unit 2 is not maintained at each inspection but is replaced after Nth inspection, and unit 3 is not maintained and replaced. Ito and Nakagawa [6] considered the degradation effect of power on–off cycles for the hazard rate of the one shot system.

Yun et al. [21] introduced periodic inspection policies for a one shot system with two types of units. Exponential and general distributions are assumed for the failure times of type 1 units and type 2 units are replaced at pre-determined time points. They proposed two optimization models to find optimal inspection intervals over a finite time period. They derived the interval availability under the assumptions of exponential distribution for failure times of type 1 units and series structure. Therefore, the failure rate of the system is the sum of all failure rates of type 1 units, λ_s. The expected uptime in an inspection interval T is

$$\int_0^T t\lambda_s e^{-\lambda_s t} dt + Te^{-\lambda_s T} \tag{21}$$

and let t_a and t_b is the system moving times in which none of the units fail or at least one unit fails, respectively.

The expected total time of one inspection cycle is

$$T + t_a e^{-\lambda_s T} + t_b(1 - e^{-\lambda_s T}) \tag{22}$$

Thus, the interval availability $A(0, T)$, is the proportion of the expected uptime within an inspection interval T,

$$A(0,\ T) = \frac{1 - e^{-\lambda_s T}}{\lambda_s\left[T + t_a e^{-\lambda_s T} + t_b(1 - e^{-\lambda_s T})\right]} \tag{23}$$

The objective is to determine optimal inspection intervals with given replacement times of type 2 units and they use the following heuristic procedure to obtain reasonable inspection intervals to satisfy the target interval availability.

Step 1 Input data of reliability and maintainability of units and set the target interval availability, A_T.

Step 2 Generate the initial inspection schedule only with replacement times of type 2 units.

Step 3 Obtain the interval availability, A_k (for all k) for the current inspection schedule [refer to (Eq. 24)].

a. If $A_T \leq A_k$ (for all k), terminate the procedure.
b. If $A_T > A_k$ (for some k), select the longest period among inspection periods that does not satisfy the target interval availability and go to Step 3.

Step 4 Add another inspection point in the inspection period, and go back to Step 2.

4.2.2 Periodic and Sequential Inspection Policies of Multi-unit One Shot Systems

Yun et al. [21] considered also another inspection optimization problem of a multi-unit one shot system in which the failure and repair times of type 1 units follow general distributions. Additionally, they considered a preventive maintenance in which type 1 units deteriorated are maintained preventively at inspections time points. They used the interval availability and life cycle cost as the optimization criteria, which are estimated by the discrete-event simulation model, in which the age reduction factor is used to calculate the ages of type 1 units after repair and preventive maintenance at inspection times. the expected life cycle cost consists of the total inspection cost (IC) total repair cost for type 1 units (CM), total replacement cost (RM) for type 2 units, and penalty cost (P_l). The objective function is expressed as follows

$$\begin{aligned} &\text{Min } TC_l = IC + \sum_{i=1}^{N_1} CM_i + \sum_{j=1}^{N_2} RM_j + P_l \\ &\text{Subject to } A_{lk} \geq A_T \end{aligned} \tag{24}$$

where A_T is the target availability, and penalty cost of alternative l, $P_l = TC_l \times (A_T - A_{lk}) \times \omega$ (ω is weighting factor) and the penalty cost is added to the fitness value in the genetic algorithm (GA).

A heuristic method and GA are used for generating alternative solutions. The concept of the heuristic method is to update the inspection intervals obtained from the GA, that is, to extend the inspection interval until the target interval availability is satisfied.

Step 1 Input the data for simulation and set the target interval availability.
Step 2 Generate alternative solutions (inspection intervals) from GA, and calculate the corresponding interval availability and life cycle cost by simulation.
Step 3 Check whether the interval availability for each inspection interval k satisfies the target interval availability. If not, go back to Step 2. Otherwise, update the inspection interval by a heuristic method.
Step 4 Generate statistics from the global best solution.

4.2.3 Sequential Inspection Policy with Preventive Maintenance

The one shot system considered in Yun et al. [26] is the one shot device introduced in Sect. 2, which is composed of two types of units, as shown in Fig. 4. Type 1 units fail at a random time, and a general distribution model is used to describe their failure time. Because type 1 units, such as the guidance system assembly in missile, are intended to age over time [47]. A Weibull distribution with an increasing failure rate is generally used as its failure time distribution. The failures of type 1 units can be detected and maintained at inspection. Type 2 units, such as the main warhead in missiles, degrade with time, and defects of type 2 units grow according to deterioration processes modeled as gamma processes. The failures of type 2 units only can be identified by destructive inspection. Therefore, an age-based preventive maintenance policy for type 1 units and a condition-based preventive maintenance policy [56] for type 2 units are applied to improve the system performance measure.

Inspection with maintenance policies proposed are summarized as follows:

- The one shot system is inspected at time t_p ($p = 1, 2,\ldots$).
- Corrective replacement is performed when failures of type 1 units are detected at inspection or the degradation level of type 2 units exceeds a pre-determined failure threshold.
- A preventive replacement is performed when the age of type 1 units exceeds a preventive replacement age, or degradation level of a defect exceeds a preventive maintenance threshold at inspection.

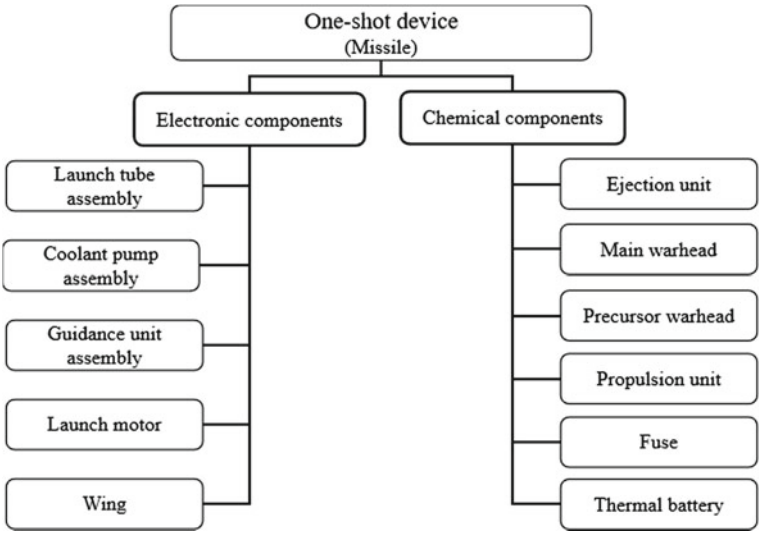


Fig. 4 A one shot system in Yun et al. [26]

Life cycle cost (LC) is used as an optimization criterion to evaluate the system performance, including the inspection operation cost (C_I) of the system and corrective maintenance ($C^{\mathrm{I}}{}_{CM}$) and preventive maintenance ($C^{\mathrm{I}}{}_{PM}$) costs of each unit i ($i = 1, 2,\ldots, N$) in the one shot system based on the maintenance actions. Interval availability is also included, defined as a mean proportion of uptimes within the inspection intervals, as another optimization criterion. They aimed to simultaneously determine the optimal inspection interval, preventive replacement ages of type 1 units, and PM thresholds of type 2 units in the one shot system that satisfies the target interval availability (A_T) and minimizes the total life cycle cost.

The objective function for a one shot system is formated as :

$$\begin{aligned} &\text{Min } E[LC] = C_I \times TI + \\ &\sum_{i=1}^{N} \left\{ \left(C^i_{CM} \times E\left[CM_i\right] \right) + \left(C^i_{PM} \times E\left[PM_i\right] \right) \right\} \\ &subject\ to\ AI_p \geq AI_T, \forall p \end{aligned} \tag{25}$$

where TI is the total number of inspections and AI_P is the interval availability of pth period.

To solve this optimization inspection model, they proposed a method to optimally determine inspection intervals of a one shot system. First, a simulation-based optimization procedure using a hybrid genetic algorithm is proposed. A gamma-bridge sampling method [57] is applied to simulate the gamma process for describing the degradation process of type 2 units. Second, they explained a hybrid genetic algorithm with a heuristic method to generate alternatives. In the hybrid genetic algorithm, they first generated the decision variables using a genetic algorithm, and then the heuristic method is used to improve the interval availability by adjusting the generated decision variables. Finally, the adjusted decision variables are checked again.

Similar to the optimization approach of Yun et al. [21], a hybrid genetic algorithm with a heuristic method is used to solve the problem. The detailed procedure is as follows:

Step 1 Input simulation data and set target availability.

Step 2 Generate alternatives using the hybrid genetic algorithm with a heuristic method.

a. Generate inspection intervals, preventive replacement ages of type 1 units, and PM thresholds of type 2 units by the genetic algorithm.
b. Improve the interval availability of alternatives by the heuristic method.

Step 3 Estimate the interval availability and life cycle cost through simulation.

a. If $AI_p \geq A_T$ (for all p), set the current best solution as the global best solution and terminate the procedure.
b. If $AI_p \geq A_T$ (for at least one period), go back to Step 2 to generate new alternatives.

5 Conclusions and Further Studies

Over the past decades, the importance of modeling one shot systems for system reliability and inspection optimization policies has been recognized by researchers and field engineers. We summarized the existing literature on system structure, unit failure mode characteristics, operation environment, and other descriptions. The storage reliability quantification methods and the optimization inspection strategies under different inspection and maintenance policies are classified and reviewed.

Martinez [45] proposed firstly a storage reliability throughout the life cycle of storage devices from the perspective of inspection, to consider the inspection error from the factory inspection to the storage inspection cycle. Several papers continuously provided more rigorous analysis and improvements of the mathematical model in Martinez [45], including the refinement of the system from the single unit to multi units, and the generalization of the failure time distribution of the units in the reliability model. Since the one-shot systems are maintained after inspection, the systems are included to repairable systems and the instantaneous or mean availabilities as the performance evaluation criteria of the system are more suitable. Yun et al. [21] proposed and defined the interval availability between inspections since the missions of one shot system can occur suddenly between inspections, which is more suitable to evaluate the performance of a system subjected to continuous inspection and maintenance activities during the storage period. Then, the interval availability model for a multi-unit one shot system when the inspection is imperfect and periodic was summarized. Later, the interval availability model of multi-unit one-shot systems under periodic and imperfect inspection [5] and the discrete-event simulation model considering a more complex operation, inspection, and maintenance processes with support equipment were discussed respectively.

Through classification analysis, we found that availability and cost terms are mainly used as optimization criteria in the research on inspection optimization problems. The common inspection optimization inspection models for one shot systems include (1) determine the inspection and maintenance policies (inspection interval/inspection time point/overhaul time point, etc.) to maximize availability(instantaneous, mean, or interval availability); (2) determine the minimum costs (expected cost rate, expected cost over the renewal cycle or life cycle cost) of the inspection and maintenance plan; (3) determine the inspection and maintenance plan that satisfies the cost constraint and maximizes the availability; (4) determine the inspection and maintenance plan that satisfies the availability constraint and minimizes the cost. Based on the earlier works on optimization models and unit characteristics analysis, Yun et al. [26] further considered the age-based PM for type 1 units, which age over time, and the condition-based PM for type 2 units, which degrades with time. They determined the optimal inspection times, the optimal preventive maintenance age for type 1 units, and the optimal preventive maintenance threshold for type 2 units simultaneously by simulation.

With the increasing scale and more complex inspection and maintenance policies of one shot systems, the interval availability and life cycle cost are hard to

be obtained by an analytical method. Instead, Monte Carlo and discrete-event simulations have frequently appeared in the recent literature as system availability evaluation methods. Several meta-heuristic methods are designed to find the optimal solutions for inspection optimization problems. Even though the proposed approach is capable of addressing complex optimization problems of one-shot systems involving realistic models and general failure distributions, more efficient simulation and optimization algorithm are required and the analytical models and optimization topics are still promising research area.

The existing research results of storage reliability and inspection models of one-shot systems this topic can be extended to the modeling and maintenance optimization of the availability of hidden failures, stand-by structure, protection systems and safety systems with the nature of "the time of failure is unknown and the system state can only be identified by inspection".

Based on the above review, the following ideas can be used as the research direction or topics for further research.

- One shot systems usually have batch and long-term storage modes. Therefore, the more scientific sampling method and the availability model of one-shot systems under different sampling inspection methods are worth for further discussion.
- The repair network (operation, inspection, and maintenance) of one shot systems like missiles is a multi-echelon maintenance system. Under different storage plans, the inspection policy and corresponding maintenance actions are also a research direction that needs further exploration
- More practical unit failure models and mechanisms could be discussed; Multiple failure modes of electronic units, such as intermittent or cycle type failures, cascading failures, coupling effect between electronic and chemical units, and interaction between failure and inspection.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant No. 72101240), the China Postdoctoral Science Foundation (Grant No. 2022M712860), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) and funded by the Ministry of Education, Science, and Technology (2022R1F1A1071057116218206530010 1) The authors give the special thanks to Professor Huang who gave them a lot of chances to present their papers and exchange their idea with many reliability researchers by organizing QR2MSE Conferences.

References

1. Cui L, Zhao X, Shen J, Xu Y (2010) An availability model for storage products under periodical inspections. Int J Reliab Qual Saf Eng 17(02):89–103
2. Kim HW, Yun WY (2016) Reliability analysis for one-shot systems with periodic inspection. J Korean Inst Ind Eng 42(1):20–29
3. Kitagawa T, Yuge T, Yanagi S (2016) Periodic and non-periodic inspection policies for a one-shot system with minimal repair. J Jpn Ind Manage Assoc 66(4E):387–395

4. Kitagawa T, Yuge T, Yanagi S (2018) Replacement timing for a one-shot system with minimal repair. Malaya J Matematik (MJM) (1, 2018):84–89
5. Zhao QQ, Yun WY (2019) Storage availability of one-shot system under periodic inspection considering inspection error. Reliab Eng Syst Saf 186:120–133
6. Ito K, Nakagawa T (2000) Optimal inspection policies for a storage system with degradation at periodic tests. Math Comput Model 31(10–12):191–195
7. Ito K, Nakagawa T (1992) Optimal inspection policies for a system in storage. Comput Math Appl 24(1–2):87–90
8. Ito K, Nakagawa T, Nishi K (1995) Extended optimal inspection policies for a system in storage. Math Comput Model 22(10–12):83–87
9. Kitagawa T, Yuge T, Yanagi S (2015) Optimum maintenance policy for a one-shot system with series structure considering minimal repair. Appl Math 6(02):326
10. Cho YS, Lee JH (2008) Optimal two-stage periodic inspection policy for maintaining storage reliability. Commun Stat Appl Methods 15(3):387–402
11. Ito K, Nakagawa T (1995) An optimal inspection policy for a storage system with three types of hazard rate functions. J Oper Res Soc Jpn 38(4):423–431
12. Ito K, Nakagawa T (1995) An optimal inspection policy for a storage system with high reliability. Microelectron Reliab 35(6):875–882
13. Nakagawa T, Mizutani S, Chen M (2010) A summary of periodic and random inspection policies. Reliab Eng Syst Saf 95(8):906–911
14. Ten Wolde M, Ghobbar AA (2013) Optimizing inspection intervals—Reliability and availability in terms of a cost model: a case study on railway carriers. Reliab Eng Syst Saf 114:137–147
15. Golmakani HR, Moakedi H (2012) Optimal non-periodic inspection scheme for a multi-component repairable system using A∗ search algorithm. Comput Ind Eng 63(4):1038–1047
16. Gen M, Cheng R, Lin L (2008) Network models and optimization: Multiobjective genetic algorithm approach. Springer
17. Liu L, Han YJ, Kim HW, Yun WY (2018) Optimal inspection intervals for multi-one-shot systems. In: Modeling and simulation in industrial engineering. Springer, Cham pp 57–83
18. Zhao QQ, Yun WY (2018) Determining the inspection intervals for one-shot systems with support equipment. Reliab Eng Syst Saf 169:63–75. https://doi.org/10.1016/j.ress.2017.08.007
19. Cheng Y, Elsayed EA (2018) Reliability modeling and optimization of operational use of one-shot units. Reliab Eng Syst Saf 176:27–36
20. Xu Y, Liao H (2015) Reliability analysis and redundancy allocation for a one-shot system containing multifunctional components. IEEE Trans Reliab 65(2):1045–1057
21. Yun WY, Han YJ, Kim HW (2014) Simulation-based inspection policies for a one-shot system in storage over a finite time span. Commun Stat Simul Comput 43(8):1979–2003
22. Jang HJ, Son YK (2011) Development of a storage-reliability estimation method using quantal response data for one-shot systems with low reliability-decreasing rates. Transactions of the Korean Society of Mechanical Engineers A 35(10):1291–1298. https://doi.org/10.3795/KSME-A.2011.35.10.1291
23. Rooney JP (1989) Storage reliability. In: Paper presented at the annual reliability and maintainability symposium, Atlanta, GA, 24–26 Jan 1989
24. Yun WY, Liu L, Han YJ, Rhee DW, Han CG (2013) Optimal inspection schedules for one-shot systems with two types of units. In: Paper presented at The international conference on quality, reliability, risk, maintenance, and safety engineering, Chengdu, 15–18 July 2013
25. van Noortwijk JM (2009) A survey of the application of gamma processes in maintenance. Reliab Eng Syst Saf 94(1):2–21
26. Yun WY, Liu L, Han YJ (2014) Metaheuristic-based inspection policy for a one-shot system with two types of units. J Mech Sci Technol 28(10):3947–3955
27. Li W, Pham H (2005) An inspection-maintenance model for systems with multiple competing processes. IEEE Trans Reliab 54(2):318–327
28. Zhang Y, Sun Y, Zhao M (2017) A combinatorial estimation approach for storage reliability with initial failures based on periodic testing data. Commun Stat Simul Comput 46(4):3319–3340

29. Merren G (1981) Dormant storage reliability assessments-data based. IEEE Trans Compon Hybrids Manuf Technol 4(4):446–454
30. Zhao M, Xie M (1994) A model of storage reliability with possible initial failures. Reliab Eng Syst Saf 43(3):269–273
31. Zhao M, Xie M, Zhang YT (1995) A study of a storage reliability estimation problem. Qual Reliab Eng Int 11(2):123–127
32. Su C, Zhang YJ, Cao BX (2012) Forecast model for real time reliability of storage system based on periodic inspection and maintenance data. Eksploatacja i Niezawodność 14:342–348
33. Zhang Y, Zhao M, Zhang S, Wang J, Zhang Y (2017) An integrated approach to estimate storage reliability with initial failures based on E-Bayesian estimates. Reliab Eng Syst Saf 159:24–36
34. Liu Z, Liu X (2018) Storage reliability assessment for the stored equipment under periodical inspection. Adv Mech Eng 10(6):1687814018782312. https://doi.org/10.1177/1687814018782312
35. Balakrishnan N, Ling MH (2014) Gamma lifetimes and one-shot device testing analysis. Reliab Eng Syst Saf 126:54–64
36. Zhu X, Liu K, He M, Balakrishnan N (2021) Reliability estimation for one-shot devices under cyclic accelerated life-testing. Reliab Eng Syst Saf 212:107595
37. Balakrishnan N, So HY, Ling MH (2015) A Bayesian approach for one-shot device testing with exponential lifetimes under competing risks. IEEE Trans Reliab 65(1):469–485
38. Azimian M, Karbasian M, Atashgar K (2021) Selecting optimal preventive maintenance periods for one-shot devices: a new fuzzy decision approach. J Prod Oper Manage 12(4):21–39
39. Leung FKN (2001) Inspection schedules when the lifetime distribution of a single-unit system is completely unknown. Eur J Oper Res 132(1):106–115
40. Maintenance theory of reliability. Springer-Verlag London
41. Sarkar J, Sarkar S (2000) Availability of a periodically inspected system under perfect repair. J Stat Plan Inference 91(1):77–90
42. Cui L, Xie M (2005) Availability of a periodically inspected system with random repair or replacement times. J Stat Plann Inference 131(1):89–100
43. Tang T, Lin D, Banjevic D, Jardine AK (2013) Availability of a system subject to hidden failure inspected at constant intervals with non-negligible downtime due to inspection and downtime due to repair/replacement. J Stat Plan Inference 143(1):176–185
44. Qiu Q, Cui L, Gao H (2017) Availability and maintenance modelling for systems subject to multiple failure modes. Comput Ind Eng 108:192–198
45. Martinez EC (1984) Storage reliability with periodic test. In: Paper presented at annual reliability and maintainability symposium, San Francisco, CA, Jan 1984, pp 24–26
46. Nelson W (1990) Accelerated testing; statistical models, test plans, and data analyses. Wiley
47. Menke JT (1983) Deterioration of electronics in storage. In: Paper presented at the National SAMPE symposium, Anaheim, California, 12–14 Apr 1983
48. Son YK, Kwon T (2016) Storage reliability estimation of one-shot systems using accelerated destructive degradation data. J Mech Sci Technol 30(10):4439–4442
49. Zequeira RI, Bérenguer C (2006) Optimal scheduling of non-perfect inspections. IMA J Manag Math 17(2):187–207
50. Zhao X, Nakagawa T (2015) Optimal periodic and random inspections with first, last and overtime policies. Int J Syst Sci 46(9):1648–1660
51. Berrade MD (2012) A two-phase inspection policy with imperfect testing. Appl Math Model 36(1):108–114
52. Nakagawa T, Mizutani S (2009) A summary of maintenance policies for a finite interval. Reliab Eng Syst Saf 94(1):89–96
53. Kaio N, Osaki S (1986) Optimal inspection policy with two types of imperfect inspection probabilities. Microelectron Reliab 26(5):935–942
54. Badia FG, Berrade MD, Campos CA (2001) Optimization of inspection intervals based on cost. J Appl Probab 38(4):872–881

55. Berrade MD, Cavalcante CA, Scarf PA (2013) Modelling imperfect inspection over a finite horizon. Reliab Eng Syst Saf 111:18–29
56. Grall A, Bérenguer C, Dieulle L (2002) A condition-based maintenance policy for stochastically deteriorating systems. Reliab Eng Syst Saf 76(2):167–180
57. Avramidis AN, Ecuyer LP, Tremblay PA (2003) Efficient simulation of gamma and variance-gamma processes. In: Winter simulation conference, vol 1, pp 319–326

Analysis for Influence of Maintenance and Manufacturing Quality on Reliability of Repairable Systems

Renyan Jiang, Wei Xue, and Yu Cao

Abstract Reliability of a repairable system is usually modeled by a failure point process with system age as underlying variable. The system may undergo a reliability improvement process due to possible technology upgrades as well as manufacturing or/and maintenance quality improvement. New methods are needed to evaluate their influence on the reliability. This chapter aims to address this issue through introducing a maintenance experience measure and concept of system technology age. They are used as the underlying variables to analyze the influence of the maintenance and manufacturing quality on the reliability, respectively. A signal-to-noise-ratio-based cluster analysis approach is also proposed to identify the change point of a function. The proposed approach can be used to determine whether or not the system undergoes a reliability growth. These concepts, approach and their appropriateness are illustrated through analyzing a real-world example that deals with a fleet of air conditioning systems of jet airplanes. The results show that the maintenance quality is poor and the fleet may undergo a reliability growth due to manufacturing quality improvement rather than due to technology upgrade.

Keywords Repairable system · Maintenance experience measure · Maintenance quality · Technology age · Manufacturing quality · Cluster analysis

1 Introduction

An important topic for a fleet of repairable systems is to build the reliability models of systems and their key components. The resulting models can be used to optimize preventive maintenance decision of the system and to forecast spare parts demand of the components [1–3]. Since the inter-failure times (IFT) of the systems depend on maintenance depth (e.g., minimal repair or overhaul) and maintenance quality, the IFT data of a repairable system are generally not independent and identically

R. Jiang (✉) · W. Xue · Y. Cao
China International Science and Technology Cooperation Base for Laser Processing Robotics, Wenzhou University, Wenzhou, China
e-mail: jiang@csust.edu.cn

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_15

distributed so that a life distribution is inappropriate as a system-level reliability model; and a failure point process is usually used as the system reliability model [4–6]. A typical model is the non-homogeneous Poisson process with a power-law mean cumulative function (MCF), whose underlying variable is system age. When the shape parameter (denoted as β) of the power-law model (PLM) is smaller than one, the system improves with its age; when $\beta > 1$, the system deteriorates; and when $\beta = 1$, the system is stationary.

The reliability improvement can be achieved through improving manufacturing quality (including technology upgrade) [7] or maintenance quality improvement resulting from the accumulation of maintenance experience [8–10]. Few works deal with the reliability analysis in such situations though the maintenance quality problems have been explored in the literature [11–13]. Thus, new analysis methods are needed to evaluate the influence of the maintenance and manufacturing quality on the reliability. This chapter aims to address this issue.

The concept of system technology age is introduced, a maintenance experience measure is defined, and a signal-to-noise-ratio-based cluster analysis approach is proposed. The maintenance quality can be evaluated through examining the short-IFT-event point process with the maintenance experience measure as the underlying variable; and the overhaul quality can be evaluated through comparing the MCFs before and after the overhaul. Similarly, the influence of manufacturing quality [technology upgrade] on reliability can be evaluated through examining the change point of the mean time between failures (MTBF) [mean time to the first failure (MTTFF)] as a function of the technology age. These concepts, approach and their appropriateness are illustrated through analyzing a real-world example that deals with a fleet of air conditioning systems of jet airplanes. The results show that the quality of both repair and overhaul is poor and the fleet may undergo a reliability growth due to manufacturing quality improvement.

The chapter is organized as follows. Section 2 deals with the technology age and maintenance experience measure, and Sect. 3 presents the non-parametric estimator of MCF and PLM. The proposed cluster analysis approach is presented in Sect. 4. Sections 5 and 6 deal with the evaluation of the maintenance quality and manufacturing quality, respectively. The chapter is concluded in Sect. 7.

2 Concepts of Different Ages

2.1 System Age, Technology Age and Maintenance Experience Measure

Consider a fleet of nominally identical repairable systems, and each system is called a unit. Referring to Fig. 1, let T_i denote the time for the i th unit ($1 \leq i \leq n$) to be put into use, w_i denote the length of the observation window of the i th unit, and $W = \max(w_i)$ denote the length of the observation window of the fleet. The origin

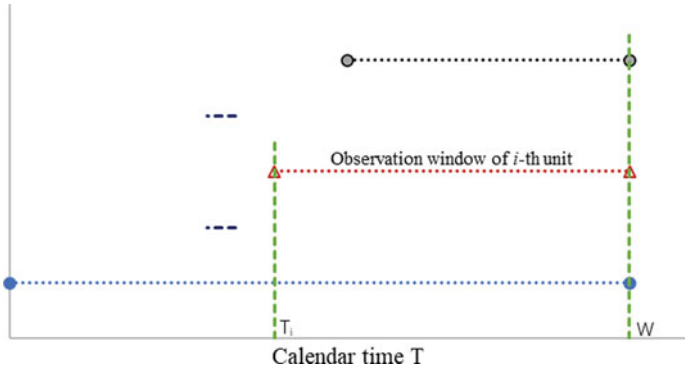


Fig. 1 Observation windows of fleet and its units

of calendar time T is defined as the left end point of the observation window of the fleet, which is the earliest time when a certain unit is put into operation. Without loss of generality, assume that the right end points of the observation windows of all the units are the same. Thus, the observation window of the i th unit can be written as (T_i, W) with $T_i = W - w_i$.

The IFTs of the ith unit are denoted as

$$(x_{ij},\ 1 \le j \le n_i) \tag{1}$$

where n_i is the total failure number of Unit i. The time to the first failure (TTFF) of unit i is x_{i1}, and the MTTFF of the fleet is given by

$$\mu_0 = \frac{1}{n}\sum_{i=1}^{n} x_{i1} \tag{2}$$

The MTBF of Unit i is given by

$$\mu_i = \frac{1}{n_i - 1}\sum_{j=2}^{n_i} x_{ij} \tag{3}$$

which excludes the influence of x_{i1}, or

$$\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{4}$$

which includes the influence of x_{i1}. The MTBF of the fleet based on (3) is defined as

$$\mu_B = \frac{1}{N_0 - n} \sum_{i=1}^{n} (n_i - 1)\mu_i \tag{5}$$

where $N_0 = \sum_{i=1}^{n} n_i$.

The age of Unit i at the jth failure is given by

$$t_{ij} = \sum_{k=1}^{j} x_{ik}. \tag{6}$$

Let $\delta_i = w_i - t_{in_i} (\delta_i \geq 0)$ denote the time between the n_i-th failure event and censoring event, which is a right-censored time. Let $s(T)$ denote the number of the units under operation at calendar time T, which is given by.

$$s(T) = \sum_{i=1}^{n} I(T > T_i). \tag{7}$$

where $I(.) = 1$ if $T > T_i$; otherwise, $I(.) = 0$.

Generally, the maintenance quality problems decrease with accumulation of maintenance experience. Jiang et al. [14] use T to describe the accumulation amount of maintenance experience. It is more reasonable to define the total number of repairs (denoted as $N(T)$) of the fleet over (0, T) as a maintenance experience measure. Its introduction allows us to examine whether the unit undergoes a reliability improvement process due to the accumulation of maintenance experience.

The manufacturing quality may get improved and the manufactured unit may undergo technology upgrades at some calendar times. Reference [14] defines T_i as the technology age of unit i. Its introduction allows us to examine whether the unit undergoes a reliability improvement process due to the technology upgrade or/and manufacturing quality improvement.

2.2 A Real-World Example

The data shown in Table 1 come from [1] and deal with the IFTs (in days) of the air conditioning systems of a fleet of Boeing jet airplanes. The data with an asterisk means that a major overhaul is completed at some time after that failure. To facilitate the analysis, it is assumed that the overhaul is conducted immediately after that failure and the time to complete the overhaul is omitted.

Since no information is available about the values of T_i and δ_i, we assume

$$\delta_i = \frac{\mu_i}{2}. \tag{8}$$

The length of the observation window of the ith unit is given by

Table 1 IFTs of air conditioning systems

Unit	x_{ij}									
1	194	15	41	29	33	181				
2	413	14	58	37	100	65	9	169	447	184
	36	201	118*	34	31	18	18	67	57	62
	7	22	34							
3	90	10	60	186	61	49	14	24	56	20
	79	84	44	59	29	118	25	156	310	76
	26	44	23	62*	130	208	70	101	208	
4	74	57	48	29	502	12	70	21	29	386
	59	27*	153	26	326					
5	55	320	56	104	220	239	47	246	176	182
	33*	15	104	35						
6	23	261	87	7	120	14	62	47	225	71
	246	21	42	20	5	12	120	11	3	14
	71	11	14	11	16	90	1	16	52	95
7	97	51	11	4	141	18	142	68	77	80
	1	16	106	206	82	54	31	216	46	111
	39	63	18	191	18	163	24			
8	50	44	102	72	22	39	3	15	197	188
	79	88	46	5	5	36	22	139	210	97
	30	23	13	14						
9	359	9	12	270	603	3	104	2	438	
10	50	254	5	283	35	12				
11	130	493								
12	487	18	100	7	98	5	85	91	43	230
	3	130								
13	102	209	14	57	54	32	67	59	134	152
	2	14	230	66	61	34				

$$w_i = \delta_i + t_{in_i} \tag{9}$$

and the time to be put into use is calculated by

$$T_i = W - w_i. \tag{10}$$

Table 2 shows the observation windows and life characteristics of units. As seen, the observation window of the fleet is determined by Unit 3 with $W = w_3 = 4665$.

From (2) yields $\mu_0 = 163.4$, and from (5) yields $\mu_B = 88.45$. These imply that the units are deteriorating with time.

Table 2 Observation windows and life characteristics of units

Unit	x_{i1}	μ_i, (3)	μ_i, (4)	δ_i	w_i	T_i
1	194	59.80	82.17	29.90	988	3677
2	413	81.27	95.70	40.64	2735	1930
3	90	83.29	83.52	41.64	4665	0
4	74	124.6	121.3	62.32	4303	361
5	55	136.7	130.9	68.35	3719	945
6	23	60.86	59.60	30.43	3650	1014
7	97	76.04	76.81	38.02	3900	765
8	50	64.74	64.13	32.37	3645	1019
9	359	180.1	200.0	90.06	3429	1236
10	50	117.8	106.5	58.90	2498	2167
11	130	493.0	311.5	246.5	1509	3156
12	487	73.64	108.1	36.82	1957	2708
13	102	79.00	80.44	39.50	2624	2041

Once T_i becomes known, the number of units under operation of the fleet at T can be calculated by (7), whose plot is shown in Fig. 2 and looks S-shaped. Similarly, it is easy to determine $N(T)$, whose plot is shown in Fig. 3 and also looks S-shaped.

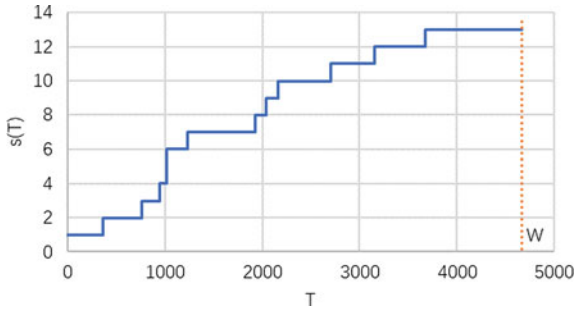


Fig. 2 Plot of $s(T)$ versus T

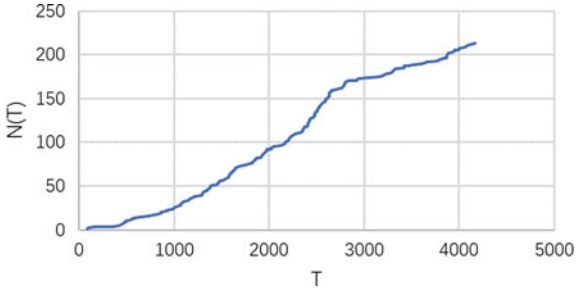


Fig. 3 Plot of maintenance experience measure as a function of calendar time

3 Mean Cumulative Function of a Failure Point Process

3.1 Non-Parametric MCF Estimator

According to (6), the failure time sequence of the i-th unit ($1 \le i \le n$) is given by (t_{ij}, $j = 1, 2, \ldots$). Ignoring the repair times, the counting process (denoted as $N_i(t)$) represents the number of failure events of the i-th unit in (0, t]. That is, $N_i(t) = j$ for $t_{ij} < t < t_{i,j+1}$. Let $N_F(t)$ denote the failure number per unit in a fleet. When the observation windows of all the units are the same, we have

$$N_F(t) = \frac{1}{n}\sum_{i=1}^{n} N_i(t). \tag{11}$$

The MCF (denoted as $M(t)$) is the expectation of $N_F(t)$, which is the mean failure number per unit in (0, t]. For a given dataset, the empirical MCF can be obtained using the Nelson estimator [15]. Specific details are as follows.

Pool the time-to-failure data (i.e., t_{ij}'s) together and sort them in an ascending order. The sorted data are denoted as

$$(t_k;\ 1 \le k \le N_0). \tag{12}$$

Let $s(t_k)$ denote the number of units at risk at time t_k. If $w_i = W$ for all the units, $s(t_k) = n$; otherwise, $s(t_k)$ decreases with k. The former [latter] case is called the single censoring [multiple censoring]. The MCF at t_k^+ is defined as

$$M\left(t_k^+\right) = M\left(t_{k-1}^+\right) + 1/s(t_k),\ t_0 = 0, M(0) = 0. \tag{13}$$

Since the Nelson estimator is a staircase function, we have $M\left(t_k^-\right) = M\left(t_{k-1}^+\right)$. We define the smoothed MCF at t_k as

$$M(t_k) = \left[M\left(t_{k-1}^+\right) + M\left(t_k^+\right)\right]/2. \tag{14}$$

The plot of the empirical MCF can have many different shapes [16]. According to the shape of the plot of the empirical MCF, one can select a proper theoretical model to approximate the empirical MCF. Let $M_0(t;\theta)$ denote the theoretical model with parameter set θ. The parameters can be estimated using the least squares method, which minimizes the sum of squared errors (SSE) given by

$$SSE = \sum_{i=1}^{N_0}[M_0(t_i;\theta) - M(t_i)]^2. \tag{15}$$

The most widely used MCF model is the PLM given by

$$M_0(t) = (t/\eta)^{\beta};\ \beta, \eta > 0. \tag{16}$$

When $\beta > 1$, it is convex and can describe the failure behavior of a deteriorating system; when $\beta < 1$ it is concave and can describe the failure behavior of an improving system; when $\beta \approx 1$, it is linear and can describe the failure behavior of a stationary system. Though it is flexible, the PLM is not always suitable to approximate an empirical MCF due to the complexity of shape of the empirical MCF [16].

3.2 Failure Intensity Function and Life Characteristics

For a given $M(t)$, the failure intensity function or the rate of occurrence of failure of the unit is defined as

$$m(t) = dM(t)/dt. \tag{17}$$

Typical shapes of $m(t)$ are increasing, decreasing and bathtub-shaped.
An instantaneous MTBF function can be defined as

$$\mu(t) = dt/dM(t) = 1/m(t). \tag{18}$$

The interval MTBF function can be defined as

$$\mu(t_0,\ t) = (t - t_0)/[M(t) - M(t_0)]. \tag{19}$$

Specifically, when $t_0 = 0$, we have

$$\mu(0,\ t) = t/M(t). \tag{20}$$

3.3 Illustration

Consider the data shown in Table 1. Figure 4 shows the plot of the empirical MCF of the fleet. As seen, the plot is inverse S-shaped; its left-hand side can be well approximated by the PLM with parameters $\beta = 1.235$ and $\eta = 150.4$; and the right-hand side can be well approximated by another PLM with $\beta = 1.358$ and $\eta = 200.2$. Applying the two PLM approximations to (18) yields the instantaneous MTBF function shown in Fig. 5. As seen, the MTBF decreases with the unit age, and hence the unit is deteriorating. Since $M(t) = 23.92$ at $t = 2074$, the interval MTBF in $(0,\ 2074)$ is 87.38, which is very close to μ_B $(= 88.45)$.

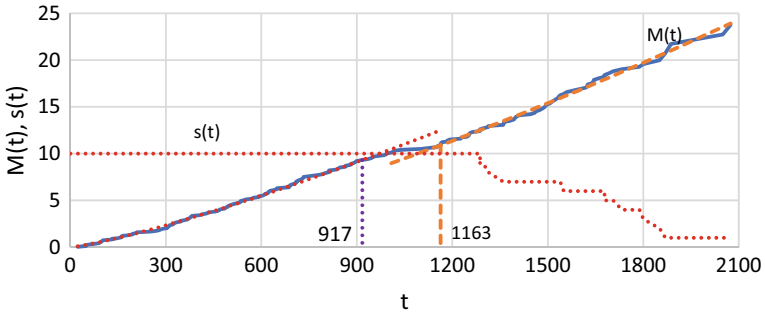


Fig. 4 Plots of $M(t)$ and $s(t)$ versus t

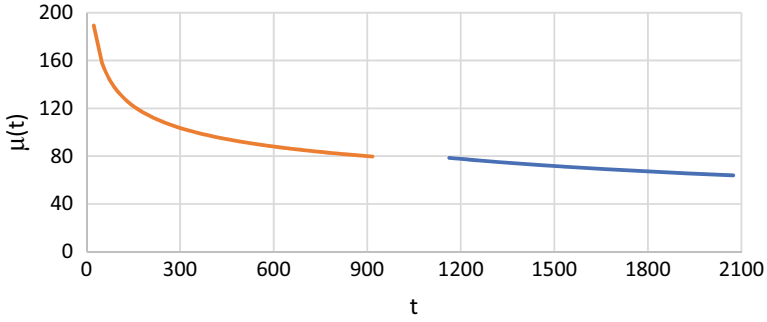


Fig. 5 Plot of $\mu(t)$ versus t

4 Cluster Analysis and Change Point

This section focuses on a signal-to-noise ratio (SNR) based cluster analysis method and its extension, which is useful for detecting the change point of a function.

4.1 Cluster Analysis for a Monotonic Dataset

Suppose that the IFT data given by (1) can be divided into two classes: small and large IFTs. The purpose of cluster analysis is to determine their boundary or critical value. The K-means approach is a widely used cluster analysis method [17]. For a one-dimensional dataset, Jiang [18] proposes a simple approach based on a similarity measure. The proposed method is applied to machine classification, group maintenance and inventory classification. For the situation where there are more than two classes, the approach will be repeatedly applied for several times. To simplify, Jiang and Huang [19] propose a SNR-based approach. Specific details are outlined as follows.

The starting point is a one-dimensional dataset like (1). Sort the data in an ascending order, and denote the sorted data as

$$(x_k; \quad 1 \le k \le K) \tag{21}$$

Then, consider an arbitrary data point, say x_p, which divides the dataset into two groups: $(x_k;\ 1 \le k \le p)$ and $(x_k;\ p+1 \le k \le K)$. The group means are calculated by

$$\mu_1 = \frac{1}{p}\sum_{k=1}^{p} x_k, \mu_2 = \frac{1}{K-p}\sum_{k=p+1}^{K} x_k. \tag{22}$$

The within-group variances are calculated by

$$v_1 = \frac{1}{p}\sum_{k=1}^{p}(x_k - \mu_1)^2, v_2 = \frac{1}{K-p}\sum_{k=p+1}^{K}(x_k - \mu_2)^2. \tag{23}$$

The between-groups distance is defined as $|\mu_2 - \mu_1|$, and the within-group distance is defined as the square root of sum of within-group variances given by $\sqrt{v_1 + v_2}$. A clustering quality measure called the SNR is defined as

$$SNR = |\mu_2 - \mu_1|/\sqrt{v_1 + v_2}. \tag{24}$$

For a good bisecting point x_p, the between-group [within-group] distance should be as large [small] as possible. As a result, the SNR should be as large as possible. Generally, the plot of SNR versus p provides intuitive information about the number of clusters and their boundaries, which corresponds to the global and local maxima of SNR. For a bisecting case, let p_0 denote the value of p that corresponds to the global maximum of the SNR.

Equation (24) does not allow $p = K$, which corresponds to the case where dataset is not classified. In this case, the SNR can be defined as

$$SNR = \mu_1/\sqrt{v_1}. \tag{25}$$

That is, the dataset does not need to be divided into two groups if $SNR|_{p=K}$ is the global maximum.

4.2 Cluster Analysis for a Non-Monotonic Dataset

In the above subsection, the dataset for clustering is not decreasing. In this subsection, we consider a two-dimensional dataset given by

$$\tau_i, y_i; 1 \le i \le K \tag{26}$$

Here, $\tau_1 \le \tau_2 \le \cdots \le \tau_K$ is a sequence of ordered time values while the dataset for clustering is $(y_i; 1 \le i \le K)$, which does not meet the following monotonical relation: $y_1 \le y_2 \le \cdots \le y_K$ or $y_1 \ge y_2 \ge \cdots \ge y_K$. Equation (26) is actually a function relation: $y_i = f(\tau_i)$. In this case, dividing the dataset $(y_i; 1 \le i \le K)$ into two group with boundary point $p = p_0$ is equivalent to find the change point of the function, which is at $\tau = \tau_{p_0}$.

The approach is the same. We still use (24) and (25) to compute the SNR for dataset $(y_i; 1 \le i \le K)$. If the plot of SNR versus p has a global maximum at $p = p_0$, the dataset given by (26) can be divided into two phases with change point τ_{p_0}: $(\tau_i, y_i; 1 \le i \le p_0)$ and $(\tau_i, y_i; p_0 + 1 \le i \le K)$. If the global maximum is at $p = K$, the function does not have the change point.

4.3 Summary

In this section, the SNR-based clustering method presented in [19] has been extended to the following two cases: the case of $p = K$ and the case of detecting the change point of a function. These will be applied in the following two sections.

5 Analysis for Influence of Maintenance Quality on Reliability

Assume that the failures before and after the overhaul are corrected by minimal repairs. Generally, the minimal repair and overhaul are carried out by different technicians while the skill of the technicians responsible for minimal repairs may get improved as $N(T)$ increases. Thus, two issues associated with the minimal repair are: (1) evaluation of minimal repair quality, and (2) to examine whether the minimal repair quality gets improved with the accumulation of maintenance experience. For those units that have experienced an overhaul, two issues are: (1) evaluation of overhaul quality, and (2) to examine whether the overhaul times are appropriate. This section addresses these issues.

5.1 Evaluation for Quality of Minimal Repairs

Among the 213 IFT data shown in Table 1, there are many data that can be thought to be small. For example, there are 20 data that are smaller than or equal to 10. A small IFT may result from poor maintenance quality. This necessitates determining a

critical value (denoted as x_c) between "small IFT" and "large IFT". The SNR-based approach presented in Sect. 4.1 can be used to determine the value of x_c. Clearly, x_c should be much smaller than the average of all the IFT data. For the example presented in Sect. 2.2, this average is μ_x=93.02 and corresponds to $K = 144$. Therefore, we first consider the dataset $(x_i; 1 \leq i \leq 144)$. Figure 6 shows the plot of SNR versus p. As seen, the SNR achieves its maximum when $p_0 = 91$, which corresponds to $x_{91} = 44$. In this case, the proportion of the small IFT data is $P = 0.4272$. Obviously, this proportion is a too large and a further clustering for the dataset with $1 \leq p \leq 91$ is needed.

Figure 7 shows the plot of SNR versus p for the dataset $(x_i; 1 \leq i \leq 91)$. In this case, the SNR achieves its maximum when $p_0 = 65$, which corresponds to $x_{65} = 27$ and $P = 0.3052$. These imply that the quality of minimal repairs is poor.

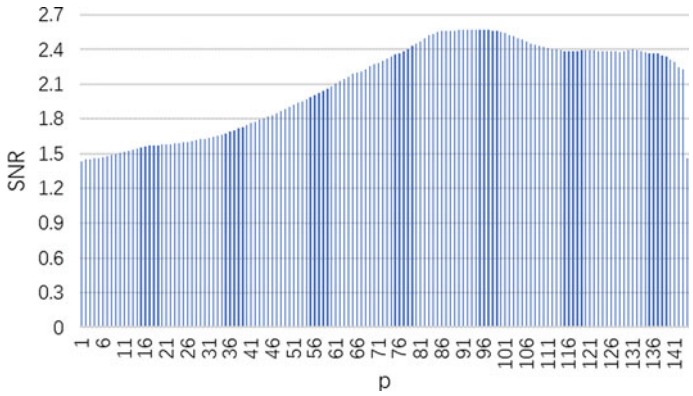


Fig. 6 Plot of SNR versus p for the data with $x_i < \mu_x$

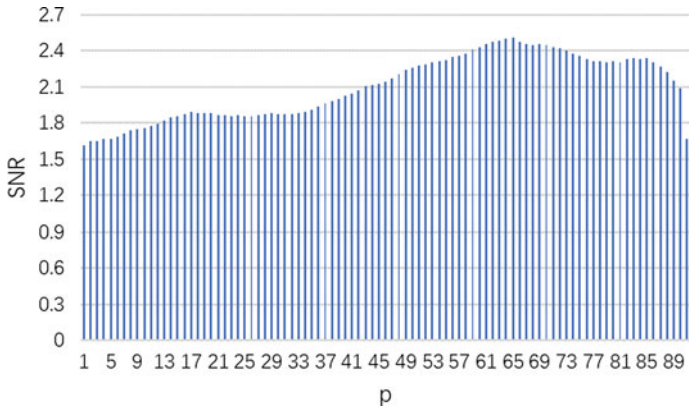


Fig. 7 Plot of SNR versus p for the dataset with $1 \leq p \leq 91$

5.2 Evaluation of Repair Quality Improvement

The maintenance quality improvement can be modeled by the point process of short IFT events with the underlying variable being the maintenance experience measure $N(T)$. This process is modeled as a marked point process, where long IFT events are marked as zero and short IFT events are marked as one [5]. Let $N_s(T)$ denote the number of short IFT events in $(0, T)$; and the event proportion is given by $P(T) = N_s(T)/N(T)$. Figure 8 shows the plot of $P(T)$ versus $N(T)$. As seen, $P(T)$ has an increasing trend, implying that no repair quality improvement occurs.

To confirm, we use the approach presented in Sect. 4.2 to find the change points of function $P(T)$. The results are displayed in Fig. 9. As seen, there are obviously two change points; one is at $p_0 = 26$ and the other is at $p_0 = 141$. They divide the range of p into three intervals: (1, 26), (27, 141) and (142, 213). The interval averages of $P(T)$ are 0.2217, 0.2863 and 0.3120, respectively. This confirms that $P(T)$ really has an increasing trend. More information is needed to explain this phenomenon.

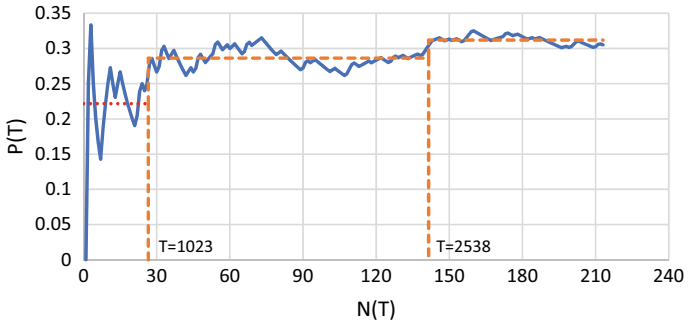


Fig. 8 Plot of $P(T)$ versus $N(T)$

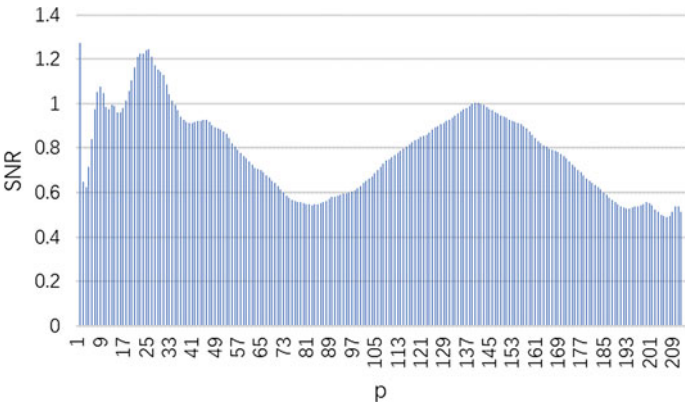


Fig. 9 Change points of $P(T)$

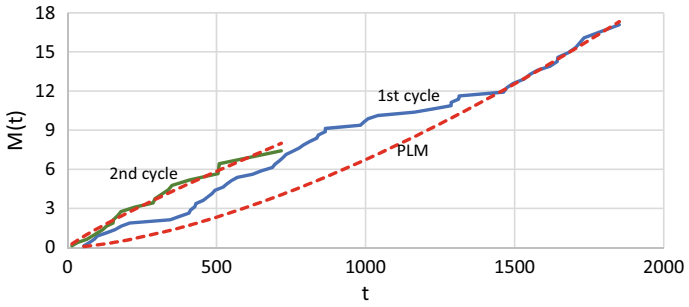


Fig. 10 Plots of MCFs of two cycles for Units 2–5

5.3 Evaluation of Overhaul Quality

Consider the data of Units 2–5 shown in Table 1. Each of them has experienced an overhaul, which divides their failure processes into two phases or two cycles: the phases before and after the overhaul. The time origin of the second cycle is the time just after the overhaul. Figure 10 shows the plots of MCFs of two cycles for Units 2–5. From the figure, we have the following observations:

- The shape of the MCF of the first cycle is roughly inversely S-shaped, and the right-hand side of the plot can be well approximated by the PLM with $\beta = 1.533$ and $\eta = 288.0$. Since $\beta \gg 1$, the MCF quickly increases after $t = 1474$, implying that an overhaul can be scheduled at some time after 1474.
- The MCF of the second cycle looks concave, and can be well approximated by the PLM with $\beta = 0.8664$ and $\eta = 65.08$. Since $\beta < 1$, the failure intensity function is decreasing, implying that the overhaul quality is poor.
- The TTFFs of the four units are 413, 90, 74 and 55, respectively, their MTTFF is 158.0; and the MTBF of the first cycle evaluated at $t = 1851$ is 108.4. On the other hand, the TTFFs after the overhauls are 34, 130, 153 and 15, respectively, their MTTFF is 83.0; and the MTBF of the second cycle evaluated at $t = 717$ is 96.7. These imply: (1) the overhaul is not as good as new and the degree of restoration is low, and (2) the overhaul quality is poor.

5.4 Optimization of Overhaul Time

The overhaul is usually expensive, and hence the overhaul time needs to be optimally determined. Let c_r and c_O denote the costs of a minimal repair and an overhaul, respectively. The cost rate function is defined in [3] as

$$J(t) = [c_O + c_r(M(t) - 1)]/t \tag{27}$$

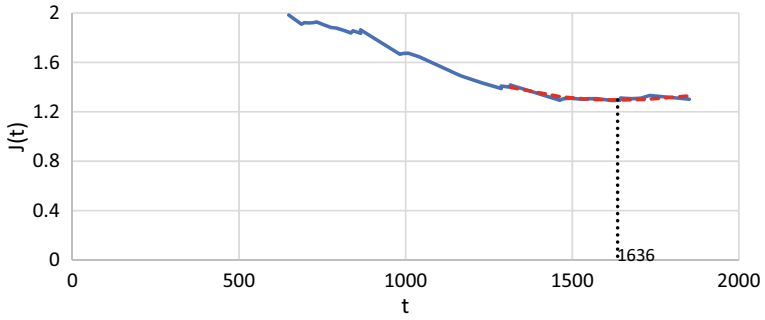


Fig. 11 Plot of cost rate function

The optimal overhaul time corresponds to the minimum of $J(t)$. When it is hard to fit the empirical MCF to a proper parametric model, (27) can be written into the following non-parametric form:

$$J(t_k) = [c_O + c_r(M(t_k) - 1)]/t_k \tag{28}$$

A curve-fitting approach can be used to fit those data points that are adjacent to the global minimum point to a regression model such as [20]

$$y(t) = a + bt + c/t \tag{29}$$

Letting $\frac{dy(t)}{dt} = 0$ yields the approximate optimal solution, given by

$$t^* = \sqrt{c/b} \tag{30}$$

For the example considered in this chapter, we take $c_r = 100$ and $c_O = 800$. Figure 11 shows the plot of cost rate function, whose global minimum point is at $t = 1620$. For the data points with $t \geq 1314$, regressing yields the coefficients of (29): $a = -2.972$, $b = 3491$ and $c = 0.001305$. From (30) yields $t^* = 1636$.

The overhaul times of Units 2–5 are at $t = 1851$, 1705, 1314 and 1678, respectively. Clearly, the overhaul times of Units 3 and 5 are fairly close to t^*, but the overhaul time of Unit 2 [Unit 4] is much larger [smaller] than t^*. For different cost ratio c_O/c_r, the conclusion can be slightly different.

5.5 Summary

According to the above analyses, we can conclude:

- For $x_c = 27$, the proportion of short IFT events reaches 30.52%, implying that the quality of minimal repairs is poor. In addition, no maintenance quality improvement occurs.
- The overhauls are imperfect, the overhaul quality is poor, and the overhaul times are not optimal.

6 Analysis for Influence of Manufacturing Quality on Reliability

In this section, we use the clustering approaches presented in Sect. 4 to examine whether the system undergoes a reliability improvement process due to technology upgrade or/and manufacturing quality improvement.

6.1 Analysis for Possible Reliability Improvement Due to Technology Upgrade

If the system undergoes a technology upgrade at certain technology age T_c, the MTBF given by (4) may have a significant increase after T_c. In other words, the dataset $(T_i, \mu_i; 1 \leq i \leq n)$ exists a change point at T_c.

For the example considered in this chapter, the values of T_i are shown in the last column of Table 2 and the values of μ_i are shown in the fourth column. Applying the approach presented in Sect. 4.2 to non-monotonical dataset $(\mu_i; 1 \leq i \leq n)$ yields the plot of SNR versus p shown in Fig. 12. As seen, the SNR achieves its global maximum at $p_0 = n = 13$, implying that the MTBF does not have a change point within the observation window of the fleet. In other words, the system has not undergone a technology upgrade.

6.2 Analysis for Possible Manufacturing Quality Improvement

The manufacturing quality can be represented by the TTFF, whose values are shown in the second column of Table 2. The TTFF as a function of T_i is shown in Fig. 13. The plot of SNR versus p associated with the TTFF dataset is shown in Fig. 14. As seen, the SNR achieves its global maximum at $p_0 = 6$, implying that the TTFF has a change point at about $T = 1020$. The average of the six [seven] MTTFs before

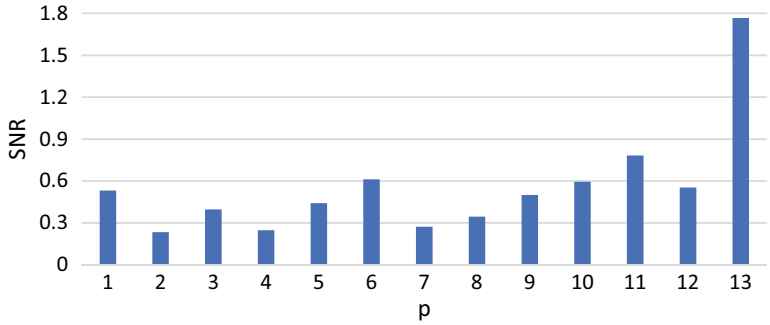


Fig. 12 Plot of *SNR* versus p associated with MTBFs given by (4)

[after] $T = 1020$ is 64.83 [247.9]. Clearly, there is a significant reliability growth after the change point, as shown in Fig. 13 (i.e., the staircase line).

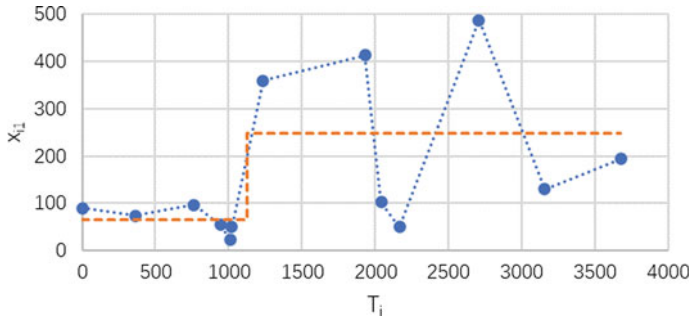


Fig. 13 Plot of x_{i1} versus T_i

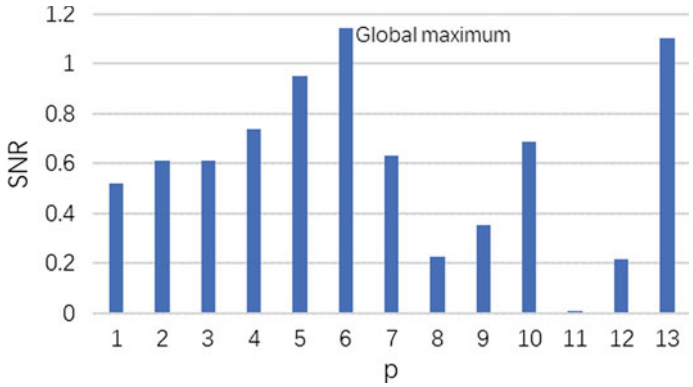


Fig. 14 Plot of SNR versus k_0 for the TTFF data

It is noted that the smallest TTFF (i.e., 23) occurs before the change point and is smaller than $x_c = 27$. Clearly, it is an early failure, which results from manufacturing quality problems. Since no technology upgrade occurs, the reliability improvement after the change point can be thought to be due to manufacturing quality improvement.

6.3 Summary

The approach to examine whether a function has a change point has been used to evaluate possible reliability improvement that may be due to technology upgrade or manufacturing quality improvement. For the example considered in this chapter, no technology upgrade occurs and there is a reliability growth due to the manufacturing quality improvement.

7 Conclusions

In this chapter, we have introduced the concepts of technology age and maintenance experience measure for a fleet of repairable systems, and the SNR-based cluster analysis method has been extended to a more general case so as to evaluate possible reliability growth due to the improvement of maintenance or/and manufacturing quality. These have been illustrated through analyzing a real-world example. The analysis found:

1. The proportion of short IFT events is high, implying that the repair quality is poor. Furthermore, no maintenance quality improvement occurs.
2. The system deteriorates with unit age, and hence the overhaul is necessary. The completed overhauls are imperfect, the overhaul quality is poor, and the overhaul times are not optimal.
3. The system has undergone a reliability growth probably due to manufacturing quality improvement rather than technology upgrade.

The concepts and analysis method presented in this chapter have a potential to evaluate the effectiveness of a reliability improvement program.

Other factors such as use conditions and environment factors affect the reliability of a repairable system. Modeling and analyzing their effect on system reliability is an open topic for future research.

Acknowledgements This research was supported in part by the Zhejiang Provincial Natural Science Foundation Projects (LZ20E050003, LD22E050001).

References

1. Proschan F (1963) Theoretical explanation of observed decreasing failure rate. Technometrics 5(3):375–383
2. Weckman GR, Shell RL, Marvel JH (2001) Modeling the reliability of repairable systems in the aviation industry. Comput Ind Eng 40(1–2):51–63
3. Jiang R (2019) Overhaul decision of repairable systems based on the power-law model fitted by a weighted least squares method. In: Asset intelligence through integration and interoperability and contemporary vibration engineering technologies. Springer, Cham, pp 277–286
4. Harold A, Feingold H (1984) Repairable systems reliability: modeling, inference, misconceptions and their causes. Marcel Dekker Inc.
5. Lindqvist BH (2006) On the statistical modeling and analysis of repairable systems. Stat Sci 21(4):532–551
6. Jiang R (2015) Introduction to quality and reliability engineering. Science Press Beijing and Springer, Berlin, Heidelberg
7. Klivans LS (1977) An improved management approach to upgrade avionic system reliability. IEEE Trans Reliab 26(1):23–28
8. Compton DW, Dunlop MD, Heim JA (1992) Improving quality through the concept of learning curves. In: Hein JA, Compton WD (eds) Manufacturing system: foundations of world-class practice. National Academy Press, Washington, DC, pp 107–115
9. Wang FK, Lee W (2001) Learning curve analysis in total productive maintenance. Omega 29(6):491–499
10. Grosse EH, Glock CH, Müller S (2015) Production economics and the learning curve: a meta-analysis. Int J Prod Econ 170:401–412
11. Duffuaa SO, Ben-Daya M (1995) Improving maintenance quality using SPC tools. J Qual Maintenance Eng
12. Scarf PA, Cavalcante CA (2012) Modelling quality in replacement and inspection maintenance. Int J Prod Econ 135(1):372–381
13. Berrade MD, Scarf PA, Cavalcante CA (2015) Some insights into the effect of maintenance quality for a protection system. IEEE Trans Reliab 64(2):661–672
14. Jiang R, Xue W, Cao Y (2022) Unit age, technology age, maintenance experience age and their application in reliability analysis of repairable systems. In: 10th Asia-Pacific international symposium on advanced reliability and maintenance, Hsinchu City, Republic of China, Oct 2022
15. Nelson W (2003) Recurrent events data analysis for product repairs, disease recurrences, and other applications. SIAM, Philadelphia
16. Jiang RENYAN, Guo YANG (2014) Estimating failure intensity of a repairable system to decide on its preventive maintenance or retirement. Int J Perform Eng 10(6):577–588
17. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31(8):651–666
18. Jiang R (2009) Cluster analysis of maintenance management problems. In: Proceedings of the IEEE international conference on industrial engineering and engineering management, Hong Kong, Dec 2009, pp 1150–1154
19. Jiang R, Huang C (2019) A signal-to-noise ratio based optimization approach for data cluster analysis. In Kim KJ, Kim H (eds) Mobile and wireless technology 2018. Lecture notes in electrical engineering, vol 513, pp 267–276
20. Jiang R, Li F, Xue W, Cao Y, Zhang K (2023) A robust mean cumulative function estimator and its application to overhaul time optimization for a fleet of heterogeneous repairable systems. Reliab Eng Syst Safety, in press

Quantification of Uncertainty of Warranty Claims

Ming Luo and Shaomin Wu

Abstract This chapter reviews the definition of warranty, introduces its different types, discusses possible causes of warranty claims, and then provides an introductory overview of the approaches to modelling warranty claims. When only warranty claim related data are available, statistical models are suggested to model the frequency of warranty claims. This approach is referred to as the black-box approach in the chapter. When the physical structure and the failure mechanism are known, both statistical models and physical models can be applied in modelling the frequency of warranty claims. This approach is referred to as the white-box approach. The chapter suggests that models that can reflect the real-world claim patterns should be the focus studied by researchers in the future.

Keywords Warranty · Point process · Uncertainty · Reliability · White-box approach · Black-box approach

1 Introduction

Warranty is a contractual obligation provided by a manufacturer to its product consumer in connection with the sale of the product. It guarantees that the product will meet certain requirements and perform its functions as specified in the warranty agreement. If the product does not meet these expectations, the manufacturer is obligated to either repair or replace the product at no charge or at a partial cost to the buyer. Warranty plays an important role in protecting consumers' interests.

There are two types of warranty: base warranty and extended warranty. Base warranty is bundled with the product at no additional cost to the buyer and extended

M. Luo (✉)
University of Bristol Business School, University of Bristol, Bristol BS8 1PY, UK
e-mail: ming.luo@bristol.ac.uk

S. Wu
Kent Business School, University of Kent, Canterbury CT2 7FS, UK

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_16

warranty is an agreement that covers repair costs after the base warranty has expired and is purchased separately and voluntarily by the buyer.

A warranty policy defines the duration of the warranty, how the product will be used, how the product will be repaired/replaced upon its failure, and how the cost of repair will be burdened. Warranty policies can be categorized into renewing or nonrenewing warranty policies. Under the renewing warranty policy, a failed product item can be replaced with a new identical item whose warranty will be renewed. Under the non-renewing warranty policy, a failed product item can either be repaired or replaced but the warranty of the repaired/replaced item cannot be renewed.

From the perspective of the length of warranty coverage, warranty can also be categorized into short term warranty and long-term warranty. Long term warranty can cover the lifetime of the product.

When a product item under warranty fails, one of the three remedy options may be offered to the buyer: a free repair/replacement, a pro rata refund, or a combination of free replacement and pro rata refund. With free replacement, the warrantor is responsible for paying the entire cost of the remedy if the product under warranty fails. Under a pro rata refund warranty, the warrantor is responsible for repair or replacement, and the cost extent of the warrantor's obligation is determined based upon the age or wear of the product at the time of failure [1]. For products sold with warranty, manufacturers bear additional cost incurred due to warranty servicing. Such warranty cost is generally substantial. For example, Volkswagen's total warranty cost in 2021 was 9.27 billion euro with 4.5% claim rate [2].

From the consumer's point of view, a warranty plays a protectional role in product purchase transactions because it provides a means of compensation if the product fails to perform as specified by the manufacturer under proper use. From the manufacturer's point of view, a warranty can be also a protection, as the warranty policy specifies the proper use and conditions of use for the product. If the product is misused, the coverage of warranty will be limited or cancelled. The warrantor is protected from the unexpected loss by the specification of requirements for care and maintenance of the product in the warranty policy [3]. Additionally, warranty also plays a promotional role in marketing a product. Many consumers believe a longer warranty indicates a more reliable product. Using warranty as a marketing tool is particularly important in promoting new products, which may be considered having a higher level of uncertainty than existing products [3]. In this sense, a warranty is an instrument which provides the consumer with a degree of assurance against uncertainty, which means the occurrence of failures of a product cannot be predicted with certainty.

A warranty does not eliminate the uncertainty of failures but transfers the burden of uncertainty from the consumer to the manufacturer during the specified period. From the manufacturer's perspective, at the strategic level, uncertainty appears at all stages from technical and commercial aspects, including design, manufacturing, marketing, etc., of the product. At the operational level, the manufacturer needs to deal with the uncertainty associated with warranty claims to achieve an efficient administration of warranty resulting in effective management of cost and reputation. Most of the warranty claims are triggered by the failures of product items; uncertainty

appears, for example, in the time to failure, when a product is put into operation. Moreover, it should also be noted that some users' behavior (i.e., human factors) may contribute to warranty claims [4], the related uncertainty needs to be considered in some cases. For example, when a failure occurs, the user can subjectively decide whether they will make a claim or not [5].

The word "uncertainty" is explained as a situation in which something is not known, or something that is not known or certain in the Cambridge Dictionary. In daily conversation, risk is another term related to this situation, and sometimes these two words are interchangeable. However, in the context of business studies, risk refers to situations under which the outcome is unknown, but the likelihood of occurrences of all potential outcomes are known, whereas uncertainty refers to situations under which either the potential outcomes and/or their probabilities of occurrences are unknown [6]. Risk can be measured by appropriate measures like variance, value-at-risk, etc., and it can be used as a handle to discuss uncertainty.

Quantifying the uncertainty of warranty claims has two tasks, (i) modeling the uncertainty and (ii) measuring the uncertainty. Quantifying the uncertainty of warranty claims is crucial for the manufacturer that offers warranty as they need to prepare resources for warranty claims. This motives our research in this chapter.

The remainder of this chapter is structured as follows. In Sect. 2, we start with reviewing the warranty claim process, identifying the uncertainties implied in different warranty claim routes, and discussing the methods of modeling the uncertainties. Following it, the methods of measuring the uncertainty are reviewed in Sect. 3. Then, this chapter is concluded in Sect. 4.

2 Warranty Claim Process and Uncertainty

The uncertainty of warranty claims has multiple sources including, but not limited to, the uncertain quality of individual products, the random operating environment, the different usage patterns, and the unknown misconception of failure of consumers. As shown in Fig. 1, the causes of warranty claims can vary. They can be due to the following causes:

(a) poor performance and management of the customer care team, e.g., poor internal training programs, poor access to product information, etc.,

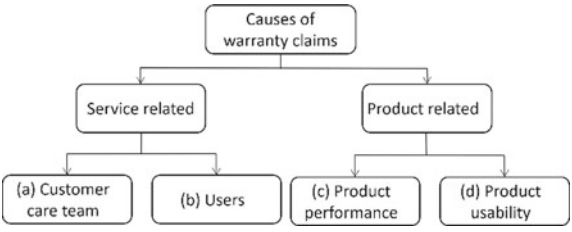


Fig. 1 Causes of warranty claims

(b) users-related, for example, abuse of product, abuse of the warranty claim process, wrong expectation of functions,
(c) product performance related, for example, hardware failure, software failure, etc., and
(d) product usability related, for example, missing accessories, poor design of product manual, etc.

The different causes of warranty claims result in different types of uncertainties, which are introduced into the warranty claim process at different stages. A today's product may be composed of a hardware subsystem and a software subsystem. Apparently, the main causes of warranty claims are related to the failures of hardware and software subsystems, on which a typical warranty claim process starts when the user believes the product is failed and results in five different routes [5], as shown in Fig. 2.

Route 1 shown in Fig. 2 describes the most typical and well-studied scenario, in which the warranty claims are triggered by the failures of the hardware subsystem and dealt with repair or replacement provided by the manufacturer. Routes 2 and 3 have been described by [4], the author indicates an absence of warranty claims is not a 'no failure' situation, it is possibly due to failed-but-not-reported (FBNR) phenomena, i.e. some users do not make warranty claims when the product items fail during

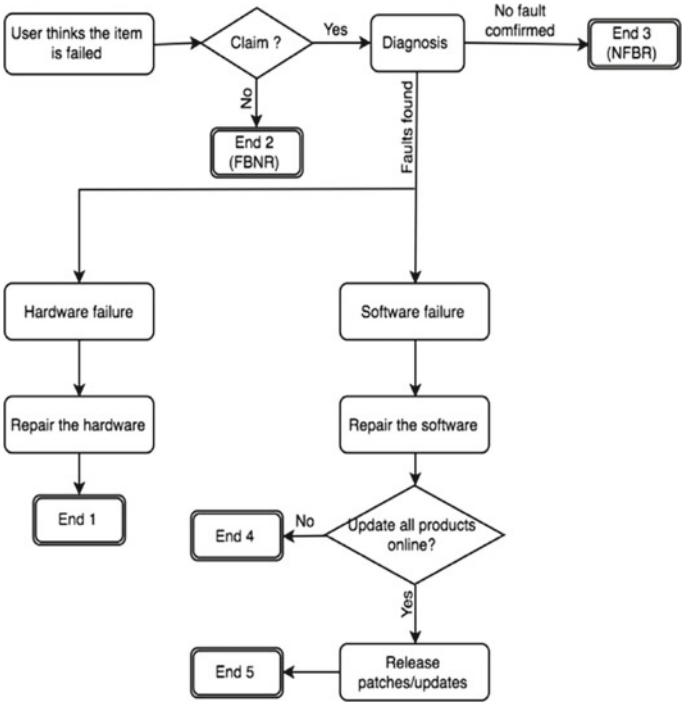


Fig. 2 Warranty claim process [5]

the warranty period; meanwhile, not all warranty claims result from product failures, because non-failed but reported (NFBR) claims can occur due to user's misconception of the product's designed functions. The FBNR and NFBR do not generate any conventional warranty claim cost but can lead to extra administration costs which should be taken into account. In these scenarios, the uncertainty is introduced by users' human behaviors. The methods of quantifying the uncertain human factors are provided in [4], which will be discussed later. Routes 4 and 5 represent the emerging scenarios in which the warranty claims result from the failures of a software subsystem. Nowadays, lots of products have a software system embedded and are connected to external data collection networks or condition monitoring networks. The complex interplays within a product and with the external environment also influence the level of uncertainty. The difference in the reliabilities of the software subsystem and the hardware subsystem should be considered in modelling warranty claims as well.

In the real world, new and used products are considered separately in developing warranty policies or conducting warranty studies. The studies of new product warranty are dominating. New products can be classified into three categories, consumer durables, industrial and commercial products for the provision of services (commercial products), and government acquisitions [8]. Most of the consumer durables are standard off-the-shelf products, while a considerable number of commercial products are custom-built or specialised. This taxonomy is used in the following discussions where it is necessary.

2.1 The Uncertainty of Warranty Claims on Hardware

Warranty claims due to the failures of the hardware subsystem are well-studied, most of the existing studies of warranty do not specify whether the research objective is due to the hardware subsystem or the software subsystem, but they are developed based on some assumptions or features of hardware subsystems, for example, the degradation of physical parts. Warranty claims are closely associated with the reliability of the hardware subsystem and influenced by the reliability of the software subsystem. In addition to the subjective decisions mentioned in Routes 2 and 3 above, several uncertain factors including usage (mode and intensity), operating environment, user skills, and maintenance also introduce uncertainty into warranty claims, as shown in Fig. 1. However, to initiate the discussion, we start with an assumption that all warranty claims are triggered by actual failures of product items, and the user makes warranty claim immediately after a failure occurred.

Overall, there are two approaches to modelling the uncertainty of warranty claims, black-box and white-box approaches. The black-box approach literally means treating the product as a black box with unseen internal details, and the uncertainty of failures is modelled without considering the mechanisms that is responsible for failure. This approach is found very useful in modeling and analyzing the warranty

claims of many products [7]. Most of the black-box approach based models are data-driven. In contrast, the white-box approach requires the knowledge of the product's internal design and the underlying mechanism of failure [7]. In most cases, a product comprises of many components, the failure of one or more critical components can lead to the failure of the entire product. Therefore, the white-box models are largely structure-related and more complex than the black-box models, but they may produce very valuable research outcomes to real world managers when modeling the warranty claims of specialized products.

2.1.1 The Black-Box Approach

In the black-box approach, products can be categorized into non-repairable and repairable. If a product is non-repairable, only data on times to first failures, i.e. times to the first warranty claims, need to be collected in modelling the survival time distribution of a product. If the product is repairable, data regarding times between claims must be collected for modelling the number of claims during the warranty coverage.

Data associated with two variables for modelling warranty claims: the time to next claim and the cost of each claim, need collecting. In case of two-dimensional warranty claim analysis, data relating age, accumulated usage, and relevant costs need collecting. The uncertainty of the time to the first claim is modeled by a continuous probability distribution defined as $F(x;\boldsymbol{\theta}) = P\{X \leq x|\boldsymbol{\theta}\}$, where x and $\boldsymbol{\theta}$ denote the time to first claim and the set of parameters of the distribution, respectively. The actual form of the distribution used in a study is decided based on the availability of data and related information. There are two distributions extensively used in studies, the Weibull distribution and the exponential distribution. The cumulative distribution function (CDF) of the Weibull distribution is defined as

$$F(x;\alpha,\beta) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}, \tag{1}$$

where α is the scale parameter and β is the shape parameter. The exponential distribution is a special case of the Weibull distribution when $\beta = 1$. Normally, the CDF of the exponential distribution is defined as

$$F(x;\lambda) = 1 - e^{-\lambda x}. \tag{2}$$

where λ is often called the rate parameter.

Estimating the cost caused by a warranty claim could be a complicated task in practice. For managerial and accounting purposes, the cost of supplies, the cost of labor, and all the cost of any related activities may be considered in the real cases. However, for the simplicity of modeling, the simplest way is to assume a constant cost per claim or a constant expected cost per claim [9]. The former ignores the uncertainty of warranty claim cost and the latter recognizes it. Some studies express

the structure of warranty claim cost as

$$C_i = A + Q_i, \tag{3}$$

where C_i is the cost of the i th claim, A is a constant, and Q_i is a random variable [10, 11]. The individual warranty claim costs of a product over a time interval can be modeled by a sequence of random variables $\{X_i, i = 1, 2, \dots\}$. When X_i are independent, the actual distribution of X_i is determined according to the information available in specific cases, and the lognormal distribution is commonly used [12].

Warranty claims may occur for many times over the warranty period of a product item, no matter whether the product is repairable or not. If the product is non-repairable, the failed product item will be replaced by a new one at the first warranty claim. The uncertainty of the time to next claim can be modeled by the same distribution of the time to first claim. The time spent on replacement is uncertain in some cases, but it could be negligible if it is relatively small compared to the warranty duration. If the product is repairable, the times between subsequent claims are influenced by the type of repair. One commonly assumed type of repair is minimal repair, which restores the failed product to the status as just before it failed, in other words, minimal repair brings the product back to a working state but does not improve the reliability of it [13].

When the repair time is negligible, the number of warranty claims on a product over the warranty period and its uncertainty can be modeled by a continuous time and discrete valued stochastic process [3], which is referred to as a point stochastic process. Such a stochastic process is denoted by $\{N(t), t \geq 0\}$ with a value space $\{0, 1, 2, \dots\}$, where $N(t)$ is the total number of points/claims in time interval $(0, t]$. If the numbers of points in disjoint time intervals, i.e. $N(t_3) - N(t_2)$ and $N(t_2) - N(t_1)$ for $t_1 < t_2 < t_3$, are independent from each other, the process has independent increment. If the numbers of point in any two time intervals of equal length, i.e. $N(t_2) - N(t_1)$ and $N(t_2 + \gamma) - N(t_1 + \gamma)$ for $t_2 > t_1$ and $\forall \gamma > 0$, have the same distribution the process has stationary increment [14].

For a non-repairable product covered by a free replacement warranty, the distributions of the time to first claim and the time between subsequent claims are independent and identical. Then, the number of claims over the warranty period can be modeled by a renewal process. If the time between claims follows an exponential distribution with CDF $F(x; \lambda) = 1 - e^{-\lambda x}$, this process is a homogeneous Poisson process (HPP), in which

$$P\{N(t) = n\} = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, n = 0, 1, 2, \dots \tag{4}$$

The expected time between claims is $\frac{1}{\lambda}$, and the expected number of claims in $(0, t]$ is λt.

For a repairable product covered by a free repair warranty, normally, minimal repairs are assumed and conducted upon warranty claims. The reliability of the hardware subsystem is changing over time, largely declining, when imperfect repairs

including minimal repairs are applied [13]. The claim arrival process is rarely regarded to be stationary. If the process has independent increment, and the time between claims follows the exponential distribution with time-varying rate parameter, the process is a non-homogeneous Poisson process (NHPP), in which

$$P\{N(t) = n\} = \frac{(\Lambda(t))^n e^{-\Lambda(t)}}{n!}; \quad n = 0, 1, 2, \ldots,$$
$$\Lambda(t) = \int_{\gamma}^{\gamma+t} \lambda(t)dt; \quad \forall \gamma \geq 0. \tag{5}$$

Here $\Lambda(t)$ is referred to as the cumulative failure intensity, and it is a positive-valued, continuous, non-decreasing function [14, 15]. $\Lambda(t)$ is the expected number of claims in $(0, t)$. In practice, the extensively used cumulative rate is defined by the power-law function,

$$\lambda(t) = \alpha\beta t^{\beta-1}; \alpha > 0, \beta > 0, \Lambda(t) = \int_{\gamma}^{\gamma+t} \lambda(t)dt = \alpha t^{\beta}. \tag{6}$$

This process is also called a Weibull process as the time to the first claim follows the Weibull distribution. It should be noted that HPP is a special case of NHPP when $\lambda(t)$ is a constant λ. Different from the HPP, the interarrival times, i.e. the time between claims, of NHPP are neither independent nor identically distributed [14].

In above discussion, the times between claims are modeled by the exponential distribution and the Weibull distribution, but in practice the distribution is not limited to these two. If we impose the times-between-claims following an identical non-negative distribution, the stochastic process becomes a renewal process [16, 17]. For instance, the ordinary renewal process (ORP) can be quoted to model the claim occurrence when a product receives perfect repairs at warranty claims and is restored to the as-good-as-new condition because of repair, and the times between claims are independent and identically distributed [17]. The expected number of claims in $(0, t]$ is defined by a renewal equation:

$$\Lambda(t) = F(t) + \int_{0}^{t} F(t - s)d\Lambda(s), \tag{7}$$

where $F(t)$ is the CDF of the time to first claim and also that of the time between subsequent claims. Obviously, the HPP is a special case of ORP, and it is one of a few special ORPs who have closed form solution of the above renewal equation [17–19].

Then, the G-renewal process, aka, generalized renewal process, relaxes the assumptions of perfect repair and minimal repair. This process provides high flexibility by assuming the effectiveness of repair to be a status between perfect repair

(i.e., replacement) and minimal repair, it can also model the effectiveness of repair to the situation of worse-than-old to better-than-new repair [17–19]. In [17], To adopt the different assumptions of repair, the repair effectiveness parameter, q, is created; and the concept virtual age is introduced

$$A_n = A_{n-1} + qX_n, n = 1, 2, \ldots \tag{8}$$

where A_n is the virtual age of the product immediately after the n th repair, $A_0 = 0$, and X_n is the time between $(n-1)$ th and n th claims. The distribution of the time to $(n+1)$ th claims after the n th repair is defined by the following conditional CDF

$$F(x|A_n = y) = \frac{F(x+y) - F(y)}{1 - F(y)}, \tag{9}$$

where $F(x)$ is the CDF of X_1, the time to first claim of a new product. The expected number of claims in $(0, t]$ is defined by the renewal equation is defined by a generalized renewal equation:

$$\Lambda(t) = \int_0^t \left(g(\tau|0) + \int_0^\tau h(x)g(\tau - x|x)dx \right) d\tau, \tag{10}$$

where $g(t|x) = \frac{f(t+qx)}{1-F(qx)}$ is the conditional probability density function (PDF) and $g(t|0) = f(t)$. When $q = 0$, the above GRP models the process under the perfect repair assumption. When $q = 1$, it models the process under the minimal repair assumption. However, due to the complexity, the closed form solution of the above generalize renewal equation has not been obtained so far. There are many different formalizations of GRP developed by researchers.

One of the popular generalizations of the renewal process is the geometric process (GP), which is used to model the monotonous stochastic point process with increasing or decreasing times between events/claims [20, 21]. For example, GP can be applied to model times-between-failures of the hardware subsystem whose successive operating times after repair are decreasing due to deterioration [20]. If the times between claims of a product is modelled by a sequence of independent nonnegative random variable $\{X_n, n = 1, 2, \ldots\}$, and the distribution of X_n is defined by a CDF $F\left(a^{n-1}x\right)$ for $a > 0$; this claim arrival process is said to be a geometric process, and a is the ratio of this GP. When $0 < a \leq 1$, this GP is stochastically increasing; and when $a \geq 1$, it is stochastically decreasing. Denote the CDF and PDF of X_1 as F and f, respectively, and set the expected value of X_1, $E[X_1] = \frac{1}{\lambda}$. Then we have

$$E[X_n] = \frac{1}{\lambda a^{n-1}}. \tag{11}$$

GP is extended to the doubly geometric process (DGP) in [21]. That is, GP becomes a special case of DGP. DGP can model a nonmonotone point process in

which the distribution of times to claim can have varying shape parameters. Based on a dataset of warranty claims on network cards, which were collected from a network card manufacturer in the UK, DGP outperforms GP in terms of the Akaike information criterion.

So far, repair is assumed to be instantaneous in the aforementioned models. However, in the real world, it is not rarely to see that the mean time to failure (MTTF) is not far longer than the mean repair time, and the time spent on repair is not negligible. The repair time can be random, and the uncertainty of this variable should be modeled in these cases as well [22].

2.1.2 The White-Box Approach

The white-box approach requires the knowledge of the internal design of a product including the details of subsystems, components, and interactions between them. In this approach, the claims or failures are modeled at the component level based on the understanding of relevant degradation mechanisms, which can be classified into two categories, over-stress and wear-out failures [3].

Over-stress failures may occur with the following physical phenomena, brittle fracture, ductile fracture, yield, buckling, large and elastic deformation, due to the degradation of material strength or/and changing stress over time. The time to failure is the first time instant when the strength falls below the stress. Wear-out failures are due to the damage of wear accumulates over time. The accumulated damage can be modeled by a stochastically increasing variable, and the failure occurs when the value of this variable reaches a threshold [3].

In recent years, warranty for complex systems or products with multiple components attract more attentions from researchers. In the daily life, lots of smart products are introduced into the market, these products, at least, have two subsystems, hardware and software. If these products are treated as single-component products or black-box systems, the important internal structure information is ignored, then, the estimation of warranty cost may be inaccurate [23], because the failure of individual components and the interaction between them can influence the reliability of product.

The configuration of the internal system structure can determine the dependence of component failures in the system. Basically, if the system has a parallel configuration, the product probably does not fail until all components fail. If the configuration is series, one component fails the product fails [23]. In reality, the system configuration can be a combination of parallel and series structures and even have a more complex structure. The interaction between components can influence the system reliability. Murthy et al. [24] describes three types of failure interactions: Type I interaction is called induced failure interaction. It assumes that the failure of a component can induce a simultaneous failure of the other components. Type II interaction is called a failure rate interaction, it assumes the failure of a component can change the failure rate of another component. Type III interaction is a combination of Types I and II.

However, a product item may be composed of multiple components at a lower level. It is not possible to model the details of all structures. To balance the efficiency

and accuracy, the individual component at a certain level can be treated as a black-box. Then, the failures of individual components will be modeled in the black-box approach, and the failures of the product will be modeled by considering the system configuration and the interaction between components. For instance, if the system configuration of a product is series, whenever a component fails the product fails. If the failed component will be replaced by a brand new one, the failure process of the component can be modeled by a renewal process. Therefore, the failure process of this product can be modeled by a superimposed renewal process [25].

2.2 The Uncertainty of Warranty Claims on Software

Generally, if the warranty claims are made immediately after the software subsystem failed, the warranty claim arrival process of the software product can be modeled by a stochastic point process like those applied on hardware products. However, the main difference between hardware failure and software failure by nature is the potential faults or causes of failures are introduced into the software during its development process. Once the fault of the software is diagnosed and removed in operation, relevant errors in codes may be debugged. These activities can result in a growing reliability of the software system. In recent years, to deal with customer's concern of software reliability, manufacturers may provide warranty on the embedded software. Similar to warranty on hardware systems, during the warranty period, the warrantor provides assurance to the customers that the software will work properly and if any defect is found, the warrantor may either repair or replace the software system without charging the customer. Different from the remedy actions conducted on hardware, the software faults can be repaired by releasing patches or updates online, and these updates can improve the reliability of all software products in the same batch [5].

2.3 Summary

The common interest of warranty management to be analysed and evaluated is the expected total warranty cost over the warranty period as well as the lifecycle of the product. This measure and the uncertainty associated summarize the financial risk or burden carried by manufacturer and even consumer [n].

3 Measuring the Uncertainty

If we come back to the discussion in the introduction section, uncertainty refers to situations under which either the potential outcomes and/or their probabilities of

occurrences are unknown. Once the uncertainty is modeled by appropriate mathematic tools, the likelihood of occurrences of all potential outcomes are known, then we can investigate the risk and give numeric measures of uncertainty [6].

Risk measures are initially introduced in financial area to meet the requirement of quantifying the losses that may be incurred. The variance of a random variable is a dominating risk measure in financial studies. However, since the variance is a symmetric risk measure, researchers turn to using downside risk measures, such as Value-at-Risk (VaR) and Conditional VaR (Expected Shortfall) to highlight the possible worst loss [27]. Furthermore, if multiple components or products are considered, using variance as a risk measure is normally applied under the assumption that the correlations between the variables of interest are linear. However, this an assumption is not imposed in the VaR and CVaR theories, potentially, VaR and CVaR can be adopted in more scenarios.

4 Conclusion

This chapter provided an introductory review on modelling the uncertainty of warranty claims. It first introduced different types of warranty claims, causes of warranty claims, and then discussed black-box and white-box approaches to modelling times between warranty claims and time to the first claims.

Our future work aims to develop models that can better reflect the real-world applications. For example, many articles in www.warrantyweek.com show that warranty claim rates are normally quite low, which implied that performing preventive maintenance policies, the focus of many warranty-related research papers, is not cost-effective. A challenge is then on the development of cost-effective approaches to reducing the cost on warranty claims.

References

1. Kapur KC, Pecht M (2014) Reliability engineering. Wiley, Hoboken, vol 86
2. Warranty Week (2022) European automaker warranty expenses, 7 Apr 2022. Warranty Week. Available at: https://www.warrantyweek.com/archive/ww20220407.html
3. Murthy DP, Blischke WR (2005) Warranty management and product manufacture. Springer Science & Business Media
4. Wu S (2011) Warranty claim analysis considering human factors. Reliab Eng Syst Saf 96(1):131–138
5. Luo M, Wu S (2019) A comprehensive analysis of warranty claims and optimal policies. Eur J Oper Res 276(1):144–159
6. Park KF, Shapira Z (2017) Risk and uncertainty. In: Augier M, Teece D (eds) The Palgrave encyclopedia of strategic management. Palgrave Macmillan, London
7. Blischke WR, Murthy DP (2011) Reliability: modeling, prediction, and optimization. Wiley
8. Murthy DNP (2006) Product warranty and reliability. Ann Oper Res 143(1):133–146
9. Kirkizoğlu Z, Karaer Ö (2022) After-sales service and warranty decisions of a durable goods manufacturer. Omega 113:102719

10. Pham H, Wang H (2001) A quasi-renewal process for software reliability and testing costs. IEEE Trans Syst Man Cybern Part A Syst Hum 31(6):623–631
11. Arnold R, Chukova S, Hayakawa Y, Marshall S (2022) Mean and variance of an alternating geometric process: an application in warranty cost analysis. Qual Reliab Eng Int 38(6):2968–2985
12. Luo M, Wu S (2018) A value-at-risk approach to optimisation of warranty policy. Eur J Oper Res 267(2):513–522
13. Luo M, Wu S, Scarf P (2022) Models of imperfect repair. In: de Almeida AT, Ekenberg L, Scarf P, Zio E, Zuo MJ (eds) Multicriteria and optimization models for risk, reliability, and maintenance decision analysis. International Series in Operations Research & Management Science. Springer, Cham, vol 321
14. Cha JH, Finkelstein M (2018) Point processes for reliability analysis: shocks and repairable systems. Springer, Cham
15. Majeske KD (2007) A non-homogeneous Poisson process predictive model for automobile warranty claims. Reliab Eng Syst Saf 92(2):243–251
16. Nakagawa T (2011) Stochastic processes: With applications to reliability theory. Springer, London
17. Kijima M, Sumita U (1986) A useful generalization of renewal theory: counting processes governed by non-negative Markovian increments. J Appl Probab 23(1):71–88
18. Kaminskiy MP, Krivtsov VV (2000) G-renewal process as a model for statistical warranty claim prediction. In: Annual reliability and maintainability symposium. 2000 Proceedings. International symposium on product quality and integrity (Cat. No. 00CH37055). IEEE, pp 276–280
19. González-Prida V, Barberá L, Márquez AC, Fernández JG (2015) Modelling the repair warranty of an industrial asset using a non-homogeneous Poisson process and a general renewal process. IMA J Manag Math 26(2):171–183
20. Lam Y (2007) The geometric process and its applications. World Scientific, Singapore
21. Wu S (2017) Doubly geometric processes and applications. J Oper Res Soc 1–13
22. Liu P, Wang G (2022) Minimal repair models with non-negligible repair time. Reliab Eng Syst Saf 217:108046
23. Liu B, Wu J, Xie M (2015) Cost analysis for multi-component system with failure interaction under renewing free-replacement warranty. Eur J Oper Res 243(3):874–882
24. Murthy DNP, Nguyen DG (1985) Study of two-component system with failure interaction. Naval Res Logist Q 32(2):239–247
25. Wu S (2019) Superimposed renewal processes in reliability. Wiley StatsRef: Statistics Reference Online, pp 1–12
26. Chukova S, Hayakawa Y (2005) Warranty cost analysis: quasi-renewal inter-repair times. Int J Q Reliab Manag
27. Babaei S, Sepehri MM, Babaei E (2015) Multi-objective portfolio optimization considering the dependence structure of asset returns. Eur J Oper Res 244(2):525–539

Prognostics and Health Management

Manufacturing Paradigm-Oriented PHM Methodologies for Cyber-Physical Systems

Ershun Pan, Tangbin Xia, and Lifeng Xi

Abstract In today's competitive environment of Industry 4.0, cyber-physical systems (CPS) of various advanced manufacturing paradigms have brought new challenges to maintenance managements. Efficient prognostics and health management (PHM) policies, which can integrate both individual machine deteriorations and different manufacturing paradigms, are urgently needed. Newly proposed PHM methodologies are systematically reviewed in this chapter: as the decision basis, an operating load based forecasting algorithm is proposed for machine health prognosis; at the machine level, a dynamic multi-attribute maintenance model is studied for diverse machines in CPS; at the system level, novel opportunistic maintenance policies are developed for complex flow-line production, mass customization and reconfigurable manufacturing systems, respectively. This framework of PHM methodologies has been validated in industrial implementations.

Keywords Maintenance · Dynamic programming · Manufacturing paradigms · Cyber-physical systems

1 Introduction

In the global competition and technique innovation, many manufacturing enterprises are pursuing a shift to cyber-physical systems (CPS) of advanced manufacturing paradigms [1]. In practice, complex flow-line production, mass customization and reconfigurable manufacturing paradigms have been applied to satisfy changeable customer demands and keep enterprise core competitiveness [2–4]. However, these CPS systems, machines and accessorial sensors have also become technologically more advanced, and more difficult to manage. This transformation provides

E. Pan · T. Xia (✉) · L. Xi
School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China
e-mail: xtbxtb@sjtu.edu.cn

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_17

motivation for improving maintenance methodologies. It is important to efficiently predict machine health statuses, eliminate unnecessary production breaks, achieve maintenance cost reduction and decrease systemic decision-making complexity [5, 6].

In the recent decades, numerous valuable studies have been devoted to the maintenance scheduling [7–9]. Prognostics and health management (PHM) has been crucial to keep CPS systems and their machines in good condition [10]. Cyber-physical systems usually consist of diverse machines, which have different degrading processes that will finally lead to failures and interrupt the normal production [11–13]. Considering CPS characters of integrated computational and physical capabilities such as actuation, sensing and communication to physical world, PHM should provide a systematical view of the machine health prognosis, the machine-level maintenance scheduling and the system-level maintenance optimization. To develop proper PHM methodologies for advanced manufacturing paradigms, it is necessary to comprehensively consider maintenance opportunities and manufacturing characters to make maintenance schedules in a cost-effective manner. However, classical opportunistic maintenance policies are insufficient to provide feasible solutions because of complex series–parallel structures, changeable batch orders and open-ended system reconfigurations [14–17]. Thus, PHM policies that can decrease decision-making complexity, avoid breakdowns of batch production, and adapt to diverse reconfigurations are urgently needed.

PHM methodologies for advanced manufacturing paradigms are complex due to the hierarchical levels of systematical maintenance decision-making: (a) accurate machine health prediction at the physical level; (b) dynamic maintenance scheduling at the machine level; (c) effective opportunistic maintenance policies at the system level. In a CPS system, recent advances in sensing and information technologies enable enterprises to on-line collect, store and process information that characterizes machine health statuses [18]. Thus, these statuses are utilized to predict machine deteriorations for supporting PHM decision-making. Furthermore, designed information transfer between the machine level and the system level should not be a "push" process, but a "pull" process. By pulling machine-level outputs, this interactive scheduling mode promotes opportunistic maintenance policies to dynamically optimize system-level schedules by integrating maintenance opportunities and manufacturing paradigms.

The remainder of this chapter is organized as follows: Sect. 2 presents a systematical PHM framework for advanced manufacturing paradigms. Section 3 proposes the W-variable forecasted-state rolling grey model (WFRGM) by considering the effect of operating loads. Section 4 develops the multi-attribute model (MAM) by utilizing the multiple attribute value theory and imperfect maintenance. Section 5 discusses the maintenance time window (MTW) for complex flow-line production, the advance-postpone balancing (APB) for mass customization, and the reconfigurable maintenance time window (RMTW) for reconfigurable manufacturing, respectively. Finally, conclusions and perspectives are drawn in Sect. 6.

2 A Systematical Framework of PHM Methodologies

Cyber-physical systems (CPS) are defined as transformative technologies for managing interconnected systems between its physical assets and computational capabilities [19]. Recent advances in manufacturing industry have paved way for a systematical deployment of CPS, within which information from all related perspectives is closely monitored and synchronized between the physical factory floor and the cyber computational space. Moreover, by utilizing advanced information analytics, networked machines will be able to perform more efficiently, collaboratively and resiliently. Cyber-physical systems are ubiquitous in power systems, transportation networks, industrial control processes, and critical infrastructures. These systems need to operate reliably in the face of unforeseen failures [20].

For understanding the impact of CPS and the relation to the manufacturing field, Monostori et al. [21] comprehensively studied cyber-physical systems in manufacturing. This important survey can help us: (1) to identify potentially impactful articles that are related to CPS and (2) to find out how CPS has evolved with respect to problems, applications and techniques. Wang et al. [22] presented the current status and advancement of cyber-physical systems and their future research directions when applied to manufacturing. The characteristics of CPS were outlined together with those of Systems of Systems (SoS), Internet of Things (IoT), Big Data and Cloud technology. Like cloud-enabled prognosis can leverage advanced manufacturing by using data and information from across the manufacturing hierarchy [23], PHM methodologies for CPS have been designed to improve efficiency, productivity, and profitability by integrating monitored information, failure prediction, system structure and manufacturing characteristics.

In industry, modern manufacturing systems with CPS technologies could be widely used in advanced manufacturing paradigms, such as complex flow-line production, mass customization and reconfigurable manufacturing paradigms. Since machine statuses are available from sensors within the cyber computational space, PHM decisions to optimize maintenance arrangements should be made in the physical factory floor by considering different manufacturing characters. Without properly integrating the special characters of advanced manufacturing paradigms, valuable information collected by CPS technologies can achieve rapid responsiveness and cost effectiveness for modern manufacturing systems. At this point, several key issues need to be addressed in the developed PHM methodologies for CPS.

(1) Based on monitored and synchronized information, it will improves forecast accuracy by incorporating real-time influencing factors (i.e., operating load) for machine health prognosis; (2) With failure frequency predictions, it is important to accurately describe hazard rate evolutions of individual machines and model machine-level maintenance operations with multiple objectives; (3) By pulling machine-level outputs, cost-effective system schedules should be studied to avoid decision-making complexity caused by series–parallel structures for complex flow-line paradigm; (4) For mass customization paradigm, an opportunistic maintenance strategy is required to handle changeable batch orders due to customer demands

and eliminate unnecessary production breaks; (5) For reconfigurable manufacturing paradigm, real-time maintenance schedules should be made to respond rapidly to diverse open-ended reconfigurations and flexible system structures.

The designed PHM framework consists of three levels, where CPS maintenance decisions are dynamically made through the machine health prognosis, the machine-level maintenance scheduling and the system-level maintenance optimization. The hierarchical scheme is shown in Fig. 1.

- *Physical level*: Cyber-physical systems of advanced manufacturing paradigms are defined as the decision objects. With rapid innovations of monitoring techniques and sensoring tools, efficient prognostic algorithm is developed to forecast accurate machine health trends for supporting the PHM decision-making process in real time, rather than over time.
- *Machine level*: For each individual machine, preventive maintenance (PM) intervals are dynamically scheduled by considering multiple attribute value theory, imperfect maintenance assessment and sequential PM scheduling mode. If a machine fails between successive PM actions, minimal repair recovers it to the failure rate that it had when it failed.

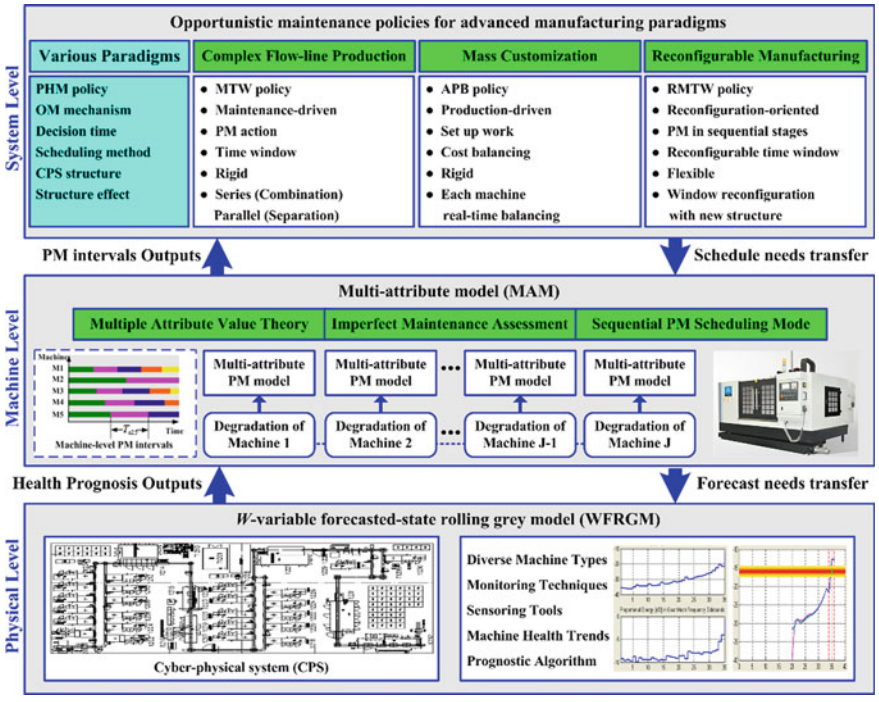


Fig. 1 Scheme of hierarchical PHM decision-making for CPS

- *System level*: By pulling PM intervals, novel opportunistic maintenance policies are presented to utilize maintenance opportunities and manufacturing characters to make dynamic maintenance schedules in a cost-effective manner. The manufacturing characters of CPS are thoroughly investigated. Thus, the proposed PHM methodologies can adapt to advanced manufacturing paradigms and achieve significant reduction of maintenance cost, production downtime and decision-making complexity.

The notation used in this chapter is listed in Table 1.

3 WFRGM Algorithm for Machine Health Prognosis

Machine health prognosis plays an important role in PHM methodologies. For complex CPS consisting of multiple machines, it is necessary to utilize maintenance opportunities and avoid production losses by forecasting machine degradations. Conventional forecasting methods can be categorized into quantitative forecasting and qualitative forecasting, including Delphi method, time series, exponential smoothing, linear regression, expert systems and neural networks [24–26]. Generally, large amounts of machine statuses are required to construct prognosis models, which limit their practical uses for CPS. In recent decades, grey model (GM) forecasting has achieved good prognosis accuracy with limited statuses by using approximate differential equations to describe future tendencies for a time series [27, 28]. The GM method, which was first proposed by Deng [29], focuses on information insufficiency and model uncertainty in analyzing future trends through studies on conditional analysis, prediction and decision making based on scarce and fuzzy information. This forecasting model is suitable for real-time prediction with limited data available.

To further increase GM accuracy, the novel philosophy comprising of utilizing practical industrial influencing factors, besides the time series itself, is needed. This study tries to achieve the following GM improvements: (1) incorporating real-time influencing factors (such as operating loads) that affect machine health trends; (2) taking new statuses into consideration and avoiding too old ones that cannot reflect current machine degradations; (3) dynamically evaluating the generating coefficient W values to overcome the shortage of static $W = 0.5$ in original GM(1,1). Thus, a W-variable Forecasted-state rolling grey model (WFRGM) is proposed to increase the accuracy of CPS health prognosis. This WFRGM algorithm includes the following steps:

(1) *Health data acquisition*: With sensing technology of CPS, health statuses of machine failure frequency at sequential time d are collected online as the in-sample testing data $x^{(0)}$, which can be represented by $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(d), ..., x^{(0)}(p))$, $p \geq 4$.

(2) *Dynamic W fitting*: In grey model, enumerate W values and select optimal ones $(W_1, W_2, W_3, ..., W_p)$ at time $d = 1, 2, 3, \ldots, p$. Evaluate the correlation coefficient (CR) of W values and corresponding operating load change rate L values

Table 1 Notation

W: Generating coefficient of grey model	L: Operating load change rate
$x^{(0)}$: Actual status of machine	$\hat{x}^{(0)}$: Forecasted status of machine
i: Index of PM cycles at machine level	j: Index of machine M_j
A_{ij}: Availability of the ith PM cycle for M_j	c_{rij}: Cost rate of the ith PM cycle for M_j
T_{pij}: Time duration of PM action	T_{fij}: Time duration of minimal repair
C_{pij}: Cost of PM action	C_{fij}: Cost of minimal repair
$\lambda_{ij}(t)$: Hazard rate function prior to the ith PM	T_{oij}: PM interval of machine level
a_{ij}: Age reduction factor	b_{ij}: Hazard increase factor
T_w: Maintenance time window	k: Index of PM cycles at system level
t_{jk}: PM time point of M_j at system level	t_k: PM execution point at system level
ETC: Excepted total system maintenance cost	c_{dj}: Downtime cost rate
u: Index of batch B_u	TB_u: Time duration of batch B_u
t_{ij}: Time point of PM from machine level	tb_u: Set-up time point after B_u at system level
$\Theta(j, tb_u)$: Maintenance decision at tb_u	G_u: PM combination set after B_u
$SCA_{j(u+1)}$: Saved cost of PM advancement	$SCP_{j(u+1)}$: Saved cost of PM postponement
$APB_{j(u+1)}$: Advance-postpone balancing	c_{sj}: Set-up cost rate
T_{pumax}: Maximum duration for PM actions	h: Index of manufacturing stage MS_h
T_{Rh}: Time duration of the hth reconfiguration	t_{Rh}: Time point of the hth reconfiguration
T_{Wh}: Time width of RMTW in MS_h	$\Theta(j, t_k)$: Maintenance decision for M_j at t_k

$(L_1, L_2, L_3, ..., L_p)$. Then construct the relationship of $W = f(L)$.

$$CR_{WL} = \frac{\sum_{d=1}^{p}\left(W_d - \overline{W}\right)\left(L_d - \overline{L}\right)}{\sqrt{\sum_{d=1}^{p}\left(W_d - \overline{W}\right)^2}\sqrt{\sum_{d=1}^{p}\left(L_d - \overline{L}\right)^2}} \tag{1}$$

(3) *WFRGM reconstruction*: With forecasted W (W_{p+1}, W_{p+2}, ..., W_{p+q}) related to real-time L (L_{p+1}, L_{p+2}, ..., L_{p+q}), WFRGM is reconstructed by taking advantages of forecasted-state rolling and generated values calculating with dynamic W in Accumulating Generation Operation (AGO).
(4) *Health trend prediction*: Then WFRGM is used to forecast the out-of-sample predictive data ($\hat{x}^{(0)}(p+1)$, $\hat{x}^{(0)}(p+2)$, $\hat{x}^{(0)}(p+3)$, ..., $\hat{x}^{(0)}(p+q)$). Forecasted-state rolling process and dynamic W values ensure a high-precision prediction, which is essential for supporting PHM scheduling. The rolling process reconstructs the grey model whenever a new status rolls in. It takes newer information into consideration and eliminates older statuses that cannot show the new machine health trend. Furthermore, in original rolling GM, the generating coefficient W is customarily given as 0.5. The static W value does not consider real-time influencing factors. Therefore, by analyzing the relationship between dynamic W values and variable L data, WFRGM can generate better forecasts.
(5) *Performance evaluation and application*: To evaluate the predicting performance, different error criteria are introduced and used, such as the mean absolute percentage error (*MAPE*) and the mean absolute error (*MAE*).

4 MAM Method for Machine-Level Maintenance Scheduling

Based on the machine health prognosis, decision makers can make maintenance schedules. With age and usage, each machine undergoes increasing wear, which finally leads to a failure and breaks the normal production. Conventional maintenance models usually suffer from a critical problem of setting periodic intervals to perform PM actions. However, it has been noticed that insufficient maintenance inevitably leads to unnecessary downtime and huge cost; on the other hand, plethoric maintenance will increase maintenance cost and decrease manufacturing profit [30]. The innovative idea of this research is to incorporate the multiple attribute value theory, the imperfect maintenance assessment and the sequential PM scheduling mode. Proper machine-level PM intervals of diverse machines will be the solid base for the opportunistic maintenance policies at the system level [31].

This research focuses on three crucial questions for optimally scheduling PM intervals: Firstly, the traditional assumption of perfect PM that covers a machine to the "as good as new" status is plausible [32]. For most machines, even though some components are replaced, the cumulative wear on adjacent components may deteriorate unnoticed. This leads to the imperfect effects of maintenance activities. In practice, a machine after PM is not as good as brand new one, that is, the hazard rate value is decreased while always greater than zero. Simultaneously, each machine tends to have more frequent maintenance since the hazard rate increases more quickly than it did in the previous PM interval. To sum up, PM not only decreases the hazard rate to a certain value but also changes the slope of the hazard rate function. Secondly, most existing maintenance models were concerning cost. In fact, it should consider

other machine-level PM objectives according to practical requirements. Thus, this study utilizes the multiple attribute value theory in building the PM model. Last but by no means the least, for responding quickly to system-level PHM pulling, a dynamic model-iteration mode is proposed to output PM intervals cycle by cycle. Since conventional static long-time planning focuses on the maintenance modelling and analysis for the whole designed lifetime and arranges PM actions in advance without considering the real-time machine degradation, which is usually not applicable in a practical factory. In our dynamic model-iteration mode, sequential PM intervals are obtained according to the real-time hazard rate evolution of the current cycles, not being relative to the whole lifetime.

The multi-attribute model (MAM), which is illustrated in Fig. 2, provides real-time PM intervals T_{oij}, even if there are L objectives (O_{1ij}, O_{2ij}, ..., O_{Lij}). The comprehensive objective function is minimized to schedule optimal PM intervals. If a smaller O_{lij} (such as the maintenance cost c_{rij}) is preferred, $\Delta_l = 0$; if a larger O_{lij} (such as the machine availability A_{ij}) is preferred, then $\Delta_k = 1$.

$$V_{ij} = w_{1ij}\frac{(-1)^{\Delta_1}O_{1ij}}{O^*_{1ij}} + w_{2ij}\frac{(-1)^{\Delta_2}O_{2ij}}{O^*_{2ij}} + \cdots + w_{Lij}\frac{(-1)^{\Delta_L}O_{Lij}}{O^*_{Lij}} \tag{2}$$

In this model, the machine availability A_{ij} and the maintenance cost rate c_{rij} may be considered as two objectives related to the efficiency and the economy, respectively:

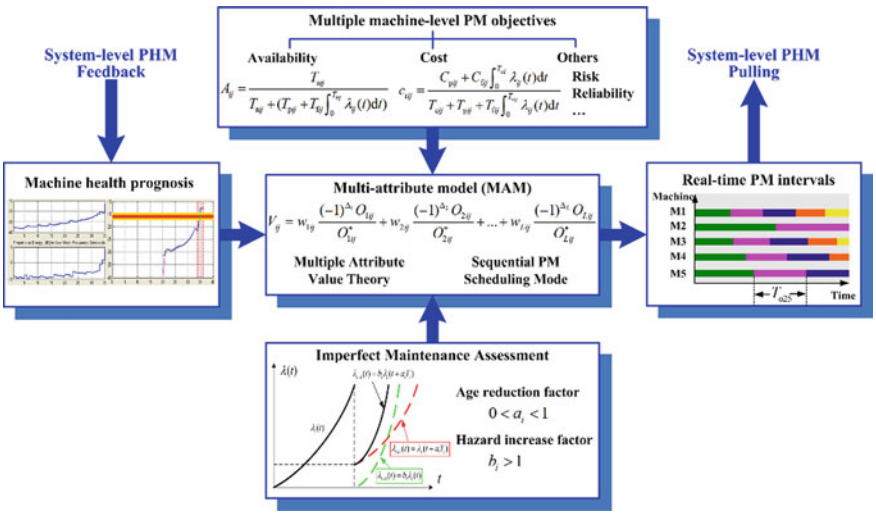


Fig. 2 Illustration of machine-level MAM method

$$A_{ij} = \frac{T_{aij}}{T_{aij} + (T_{pij} + T_{fij}\int_0^{T_{aij}} \lambda_{ij}(t)dt)} \tag{3}$$

$$c_{rij} = \frac{C_{pij} + C_{fij}\int_0^{T_{cij}} \lambda_{ij}(t)dt}{T_{cij} + (T_{pij} + T_{fij}\int_0^{T_{cij}} \lambda_{ij}(t)dt)} \tag{4}$$

For each next PM cycle, with the actual interval T_{ij} from the system-level feedback, the relationship between hazard rates of consecutive cycles can be defined as:

$$\lambda_{(i+1)j}(t) = b_{ij}\lambda_{ij}\left(t + a_{ij}T_{ij}\right), t \in (0, T_{(i+1)j}) \tag{5}$$

In imperfect maintenance effects, the age reduction factor a_{ij}, $a_{ij} \in (0, 1)$ indicates that imperfect PM causes the machine's initial failure rate to become $\lambda_{ij}(a_{ij}T_{ij})$; meanwhile, the hazard increase factor b_{ij}>1 reflects that PM increases the failure rate $b_{ij}\lambda_{ij}(t)$.

5 Opportunistic Maintenance for Various Cyber-Physical Systems

Nowadays, there has been a growing interest in PHM methodologies of multi-unit systems for leading enterprises. It is essential to investigate and model the complicated machine interactions and the diverse manufacturing characters, which provide maintenance opportunities for CPS of advanced manufacturing paradigms. Opportunistic maintenance refers to the scheme where PM can be performed at opportunities with the advantages of combining individual PM actions and saving much group maintenance cost [33, 34]. To overcome the exponential decision-making complexity with machine number increasing and apply the system-level PHM methods to advanced manufacturing paradigms, novel opportunistic maintenance policies will be presented in detail.

5.1 MTW Policy for Complex Flow-Line System

Complex series–parallel cyber-physical systems have been widely used to satisfy flow-line productions. In this article, a general PHM decision-making policy is proposed by considering both machine degradation and system structure. This maintenance time window (MTW) policy can help enterprise managers to make dynamic maintenance schedules based on not only single-machine plans, but also the whole-system global programming. MTW programming is applied by pulling real-time machine-level PM intervals. A breakdown caused by one machine is utilized to carry out PM actions on non-failed ones, thus unnecessary breakdown of CPS

could be avoided. This maintenance-driven opportunistic maintenance policy aims to systematically obtain system-level maintenance schedules in a cost-effective manner.

(1) *MTW-separation in parallel subsystem*: According to machine-level PM intervals, the MTW value T_w provides a criterion to separate PM actions in subsystems. MTW-separations can avoid the unnecessary downtime of upstream and downstream machines.
(2) *MTW-combination in series subsystem*: Pulling the outputs from MAM and MTW-separations cycle by cycle, MTW is defined as the criterion to combine PM actions within $[t_k, t_k + T_w]$. The time point t_k is when one machine is preformed PM, which also means maintenance opportunities for other machines in series.
(3) *System performance evaluation*: The total system maintenance cost (ETC) by using MAM policy can be evaluated based on system-level maintenance schedules. The total maintenance cost of the kth cycle for machine j can be evaluated by:

$$\lambda_{(i+1)j}(t) = b_{ij}\lambda_{ij}\left(t + a_{ij}T_{ij}\right), \quad t \in (0, T_{(i+1)j}) \tag{6}$$

where $\Theta(j, t_k) = 0$ means no maintenance action is initiated on Mj at the time point t_k, but this machine will be down; $\Theta(j, t_k) = 1$ means the PM action is combined to be performed in advance; $\Theta(j, t_k) = 2$ means no maintenance action is initiated and the machine continues to operate. Thus, the total system maintenance cost for the CPS in its mission lifetime can be obtained by:

$$ETC = \sum_{k=1}^{K}\left(\sum_{j=1}^{J} ETC_{kj}\right) \tag{7}$$

5.2 APB Policy for Mass Customization System

As one of advanced manufacturing paradigms, mass customization is widely used to response quickly to changeable customer demands. In mass customization, batch orders are processed through CPS with following production characteristics: (1) Batch orders are independent with diverse lot size; (2) Batches are sequentially ordered only a transient time beforehand; (3) One set-up work happens when a batch switches to another; (4) It prefers no interruptions in each batch cycle to ensure product quality. To meet the requirements of mass customization, there has been a great need to propose a new type of opportunistic maintenance that considers machine degradations and manufacturing characteristics [35].

In this study, a production-driven opportunistic maintenance policy is presented to eliminate unnecessary production breaks and achieve significant cost reduction. According to sequential batch orders and machine-level PM intervals, the advance-postpone balancing (APB) policy utilizes the set-up works and analyzes the cost savings to schedule real-time PM adjustments. Each set-up time between successive batches is used to perform PM actions, thus unnecessary breakdown during batch productions can be avoided. We apply APB programming to analyze the cost savings of PM advancement and PM postponement, and then choose the better PM adjustment to ensure no-disruptions and reduce maintenance cost.

This production-driven APB policy has the advantages for the mass customization paradigm: (1) APB satisfies no-disruption requirements during changeable batch orders, other than traditional constant production assumption; (2) It utilizes planned production downtimes as maintenance opportunities to advance or postpone PM actions, which significantly reduces the complexity of system-level scheduling; (3) By choosing the greater cost savings between PM advancement and PM postponement at each set-up opportunity, APB ensures that the maximization of ETC-saving can be achieved. The procedure of APB programming is illustrated in Fig. 3.

When each batch B_u has been finished, and the next batch B_{u+1} has not started, this moment tb_u is utilized as the decision time to schedule APB. On the one hand, if machine M_j is prevented maintained now, the saved cost by advancing PM in batch B_{u+1} can be evaluated as:

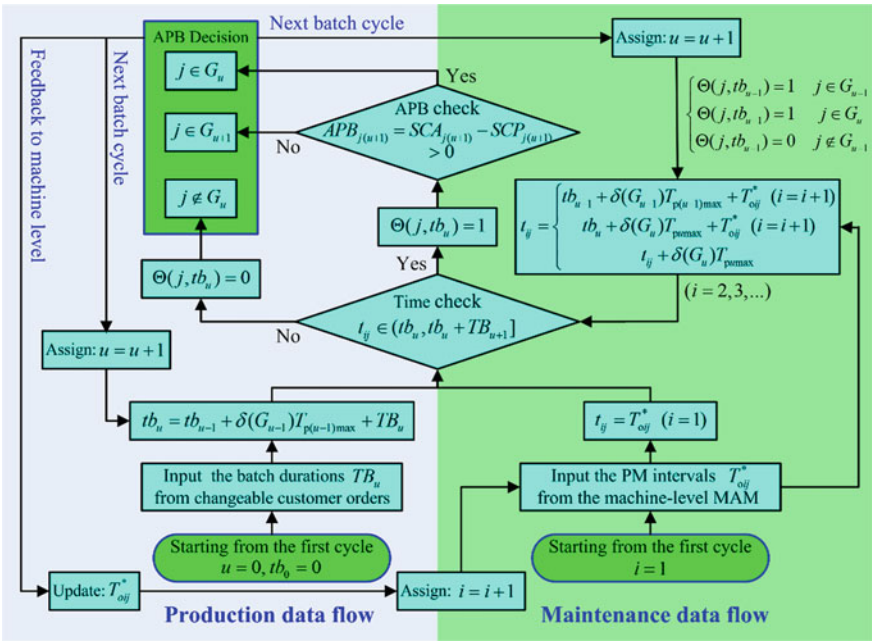


Fig. 3 Flowchart of APB policy for mass customization CPS

$$
\begin{aligned}
SCA_{j(u+1)} &= SCA^{d}_{j(u+1)} + SCA^{f}_{j(u+1)} - SCA^{p}_{j(u+1)} \\
&= T_{pij}\left(c_{dj} - c_{sj}\right) + \left[\int_{0}^{T^{*}_{oij}} \lambda_{ij}(t)dt - \int_{0}^{T^{*}_{oij}-(t_{ij}-tb_{u})} \lambda_{ij}(t)dt\right] C_{fij} \\
&\quad - \frac{t_{ij} - tb_{u}}{T^{*}_{oij} - \left(t_{ij} - tb_{u}\right)} C_{pij}
\end{aligned} \tag{8}
$$

where $SCA^{d}_{j(u+1)}$ is the downtime cost saving, $SCA^{f}_{j(u+1)}$ is the minimal repair cost saving, $SCA^{p}_{j(u+1)}$ is the PM cost saving of PM advancement.

On the other hand, if PM of machine M_j is postponed to the next set-up time point tb_{u+1}, the minimal repair cost saving will be a negative value (prolonged PM interval leads to increasing cumulative failure risk and more minimal repair cost) and the PM cost saving will be a positive value (longer intervals mean that less PM actions would be needed in the same scheduling horizon). Therefore, the saved cost by postponing PM in batch B_{u+1} can be evaluated as:

$$
\begin{aligned}
SCP_{j(u+1)} &= SCP^{d}_{j(u+1)} - SCP^{f}_{j(u+1)} + SCP^{p}_{j(u+1)} \\
&= T_{pij}\left(c_{dj} - c_{sj}\right) - \left[\int_{0}^{T^{*}_{oij}+(tb_{u+1}-t_{ij})} \lambda_{ij}(t)dt - \int_{0}^{T^{*}_{oij}} \lambda_{ij}(t)dt\right] C_{fij} \\
&\quad + \frac{tb_{u+1} - t_{ij}}{T^{*}_{oij} + \left(tb_{u+1} - t_{ij}\right)} C_{pij}
\end{aligned} \tag{9}
$$

where $SCP^{d}_{j(u+1)}$ is the downtime cost saving, $SCP^{f}_{j(u+1)}$ is the minimal repair cost saving and $SCP^{p}_{j(u+1)}$ is the PM cost saving of PM postponement.

According to the values of SCA and SCP, $APB_{j(u+1)}$ could be defined as the criterion to decide whether to advance or postpone this PM action:

$$
APB_{j(u+1)} = SCA_{j(u+1)} - SCP_{j(u+1)} \tag{10}
$$

5.3 RMTW Policy for Reconfigurable Manufacturing System

The system structure of reconfigurable manufacturing CPS can be adjusted to meet various future products and changeable market demands [36, 37]. In other words, the main advantage of reconfigurable manufacturing is the adaptability to the uncertainties of the open system architecture with reconfigurable system structures. For the entire system, those different reconfigurations are caused by the changing needs in terms of capacity and functionality, while the production process will be separated

into sequential manufacturing stages. Each manufacturing stage (MS_h) has its own system structure designed for its current production requirements. If the system-level maintenance policy has to be rebuilt according to each different structure, its responsiveness and flexibility will be obviously weakened [38].

By extending the previous research from both reconfigurable structure and manufacturing paradigm aspects, this study presents a reconfiguration-oriented opportunistic maintenance policy to achieve rapid responsiveness and cost effectiveness for future reconfigurable manufacturing. Other than rebuilding new system-level policies for different stationary structures, the developed reconfigurable maintenance time window (RMTW) focuses on the structure analysis to extract reconfigured parallel subsystems and series subsystems in each manufacturing stage. Faced with different system structures, the RMTW policy utilizes reconfiguration characters and maintenance opportunities to constantly redefine reconfiguring scheduling criteria within a uniform method. This manner is more suitable for rapidly adapting to new system structures in reconfigurable manufacturing systems (RMS).

The production scenarios in Fig. 4 can be taken as an example to illustrate the RMTW scheduling for system-level reconfigurations. After the original design, the RMS enters service at time $t = t_{R1} = 0$ with its initial system structure (5 machines). In the first manufacturing stage MS_1, the time width value of RMTW T_{W1} is defined as a criterion to separate PM actions in parallel subsystems and combine PM actions in series subsystems based on machine-level PM intervals.

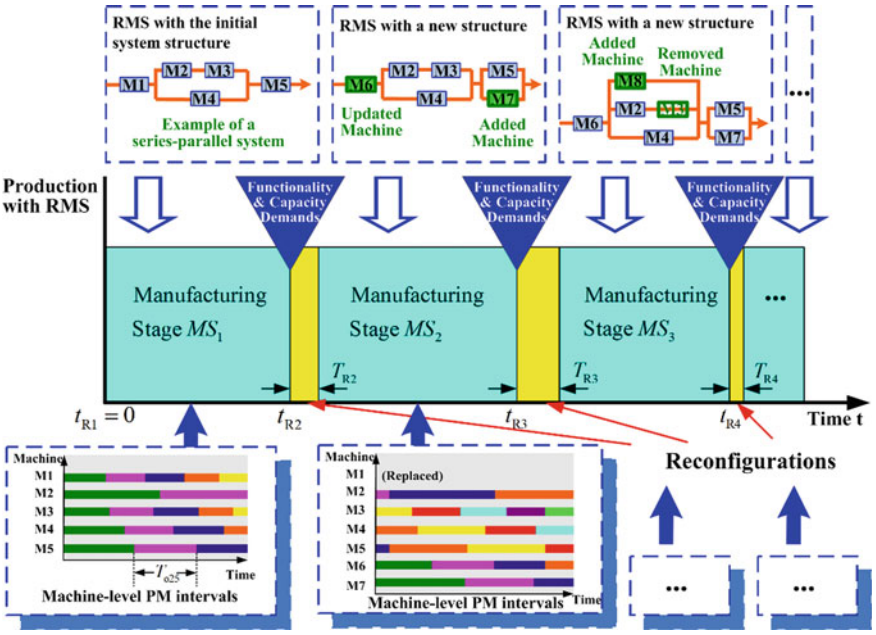


Fig. 4 Production scenarios of reconfigurable manufacturing CPS

At the reconfiguration time t_{R2}, the structure is redesigned for the second manufacturing stage MS_2. In the time duration of this reconfiguration T_{R2}, M1 is replaced with a new M6, and M7 is added in parallel with M5. Then, the RMS continues production with a new structure, while a redefined time width of RMTW T_{W2} is applied for reconfigured parallel/series subsystems to minimize the total system maintenance cost.

Similarly, in the next reconfiguration before MS_3, M3 is removed, while M8 is added in parallel with M2 and M4. In contrasted to the traditional manner of rebuilding new system-level policies for different structures, RMTW scheduling focuses on reconfiguring scheduling criteria T_{Wh} within a uniform method for rapidly adapting to new structures. Above structure analysis of each manufacturing stage is essential for RMTW scheduling. Then, the process flowchart of the proposed RMTW programming is shown in Fig. 5.

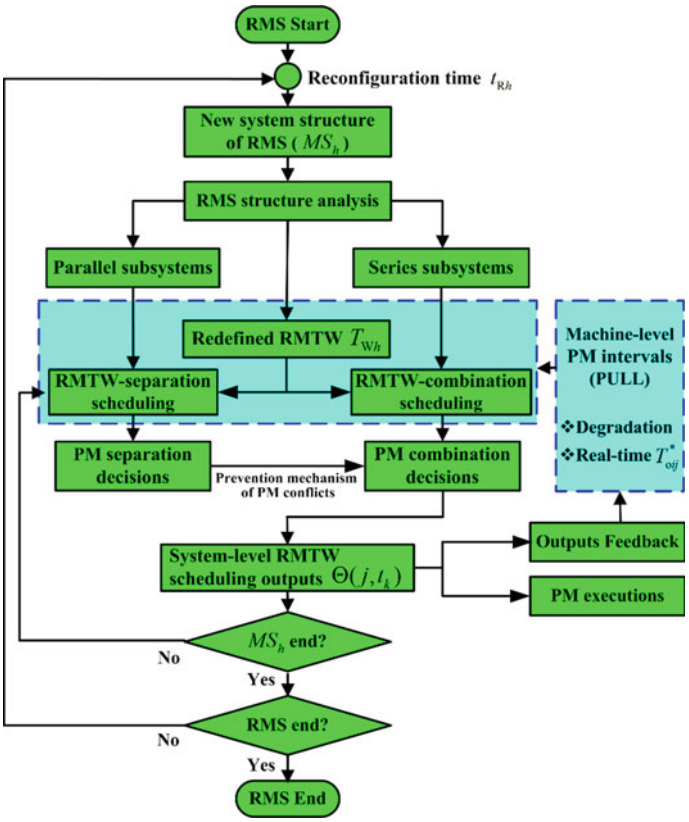


Fig. 5 Flowchart of RMTW policy for reconfigurable manufacturing CPS

6 Case Study of PHM Methodologies

6.1 Effectiveness of WFRGM Algorithm

To prove the prognosis accuracy of the proposed WFRGM algorithm, the increasing health statuses of a monitored machine's deterioration during a maintenance interval are collected. The twelve status data points of failure frequency from monitoring points 1–12 are regarded as the in-sample test data, which reflects the increasing failure risk. The remaining six states from cycles 13–18 are used for out-of-sample forecasting.

Results of the linear regression model (LRM) with $\hat{x}^{(0)}(d) = 0.0362d + 0.0211$, the original GM(1,1) model with $\hat{x}^{(0)}(d) = \left(1 - e^{-0.147}\right)(x^{(0)}(1) + 0.0951/0.147)e^{0.147(d-1)}$, the actual-state rolling grey model (ARGM), the forecasted-state rolling grey model (FRGM) and the proposed WFRGM algorithm have been presented in Table 2. The plot of actual versus forecasted machine states from above five models is shown in Fig. 6.

From the result comparisons in Table 2, it can be found that the *MAPE* (5.19%) and *MAE* (0.0395) of WFRGM are all lower than LRM (*MAPE* = 21.26%; *MAE* = 0.1608), GM (*MAPE* = 17.74%; *MAE* = 0.1415), FRGM (*MAPE* = 15.73%; *MAE* = 0.1263) and ARGM (*MAPE* = 9.80%; *MAE* = 0.0729), indicating the highly accurate forecasting ability. Thus, WFRGM algorithm can provide real-time machine health information to dynamic PHM decision-making.

6.2 Effectiveness of MAM Method

A 5-unit series–parallel system with the initial system structure in Fig. 4 is selected as an example for numerical experiments using the proposed MTW policy. In this manufacturing system, PM intervals of each machine are dynamically scheduled by the MAM method according to individual machine degradation. The reliability of each machine is formulated by a Weibull failure probability function: $\lambda_{1j}(t) = (m_j/\eta_j)(t/\eta_j)^{m_j-1}$, which has been widely used to fit repairable equipment in electronic and mechanical engineering. Machine e parameters are shown in Table 3.

From the results of industrial implementations [33], the proposed machine-level MAM method reveals following conclusions: (1) The PM interval decreases while PM cycle increases, since the underlying hazard rate evolution becomes faster with the degradation process; (2) Machine availability will be lower and maintenance cost will be higher as a machine ages due to the consideration of maintenance effects; (3) Ignoring the effects of a maintenance activity will lead to less availability and extra cost, and MAM contributes to more practicality of PM intervals.

Table 2 Forecasting results of different methods

Monitoring point	Actual status	LRM	GM	FRGM	ARGM	WFRGM
1	0.0837	0.0573	0.0837	0.0837	0.0837	0.0837
2	0.0864	0.0935	0.1157	0.1157	0.1157	0.1157
3	0.1705	0.1297	0.1340	0.1340	0.1340	0.1340
4	0.1732	0.1659	0.1552	0.1552	0.1552	0.1552
5	0.2110	0.2021	0.1798	0.1798	0.1798	0.1798
6	0.2122	0.2383	0.2082	0.2082	0.2082	0.2082
7	0.2388	0.2745	0.2412	0.2412	0.2412	0.2412
8	0.2448	0.3107	0.2794	0.2794	0.2794	0.2794
9	0.2856	0.3469	0.3236	0.3236	0.3236	0.3236
10	0.4144	0.3831	0.3749	0.3749	0.3749	0.3749
11	0.4219	0.4193	0.4342	0.4342	0.4342	0.4342
12	0.5318	0.4555	0.5030	0.5030	0.5030	0.5030
In-sample testing statuses (1–12)						
13	0.6316	0.4917	0.5827	0.5827	0.5827	0.6207
14	0.6401	0.5279	0.6749	0.6586	0.6857	0.6743
15	0.6812	0.5641	0.7818	0.7649	0.7762	0.7074
16	0.7483	0.6003	0.9056	0.8849	0.8421	0.8285
17	0.8540	0.6365	1.0490	1.0315	0.9130	0.8589
18	0.9026	0.6727	1.2151	1.1953	0.9976	0.9830
Out-of-sample forecasting statuses (13–18)						
MAPE (%)		21.26	17.74	15.73	9.80	5.19
MAE		0.1608	0.1415	0.1263	0.0729	0.0395

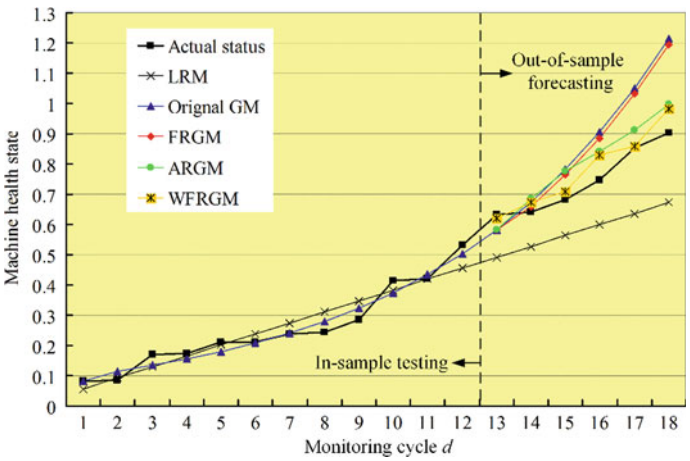


Fig. 6 Comparison of actual and forecast machine health states

Table 3 Machine parameters

j	m_j	η_j	T_{pij}	T_{fij}	C_{pij}	C_{fij}	c_{dij}	a_{ij}	b_{ij}
1	3.0	8000	140	600	5000	35,000	80	$i/(15i+5)$	$(17i+1)/(16\ i+1)$
2	2.0	7000	120	200	6000	18,000	40	0.03	1.04
3	1.5	12,000	200	350	2000	15,000	30	$i/(20\ i+20)$	1.03
4	3.0	13,000	80	300	7500	22,000	45	0.025	$(16i+3)/(15\ i+3)$
5	2.5	16,000	300	800	2500	25,000	75	$i/(16\ i+14)$	1.05

6.3 *Effectiveness of MTW Policy*

To validate the MTW policy for complex flow-line systems, we program the system-level maintenance schedule with machine parameters in Table 3. Taken $T_w = 800$ h for the MTW programming as an example, the CPS mission lifetime is 25,000 h. Table 4 provides the system-level maintenance schedule results.

The influence of MTW-value and the effectiveness of MTW programming is shown in Fig. 7. It is proven that MAM policy can reduce ETC up to 27% comparing with Individual maintenance mode (IMM) of $T_w = 0$ and Simultaneous maintenance mode (SMM) of $T_w = 25{,}000$. Besides, it can be concluded that larger MTW value enables more machines to take advantage of maintenance opportunities, but too large MTW causes extra maintenance and more ETC will be needed for CPS.

Moreover, traditional opportunistic maintenance policies calculate the cost-savings of all possible combinations at each cycle with the exponential decision-making complexity of $O(2^{J-1})$. For example, Zhou et al. [16] took a 3-unit system to illustrate the opportunistic PM scheduling algorithm, while the cost savings for 4 possible combinations were calculated at each opportunity. For our presented MTW policy, since the numbers of parallel/series subsystems and their respective machines are all smaller than J, the maximal decision-making complexity at each opportunity is less than $2J^2$. Thus, the MTW complexity is just polynomial with total machine number J, which means even a complex flow-line CPS with a large number of machines can be handled.

Table 4 System-level maintenance schedule based on MTW

j	Time point of PM activity (hours)											
1	3319		6911	10,020	13,121		15,134	17,940		20,212		22,789
2	3319		6911	10,020			15,134		19,105			22,789
3		5108		10,020			15,134			20,212		
4			6911			14,455					21,984	
5		5108		10,020			15,134			20,212		

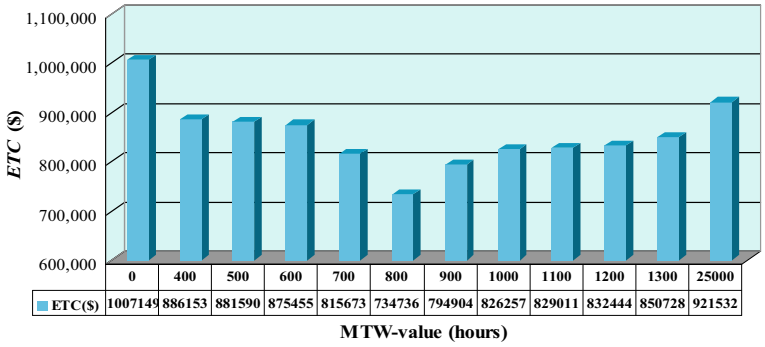


Fig. 7 ETC of the flow-line CPS with various MTW

6.4 Effectiveness of APB Policy

Faced with sequential batch orders, APB dynamically utilizes set-up works and analyzes the cost savings to reduce the total system maintenance cost. PM intervals and various batch orders are pulled to make opportunistic maintenances cycle by cycle. For a 7-unit mass customization CPS, results of production-driven opportunistic maintenance are presented in Table 5.

The results from mass customization CPS (Fig. 8) reveal that the mechanism of APB policy can ensure the lowest ETC. On the one hand, huge downtime cost saving ensures that ETC of APB policy is lower than those of maintenance-driven opportunistic maintenance policies (e.g. IMM, SMM and MTW). On the other hand, APB dynamically compares cost savings and chooses PM adjustment with $Max\{SCA_{ju}, SCP_{ju}\}$, which is thus a more cost-effective policy than Advanced maintenance mode (AMM) and Postponed maintenance mode (PMM). Therefore, APB policy achieves significant cost reduction by considering batch characteristics and making PM adjustment based on maximum cost saving for each machine at each set-up time.

Table 5 APB results in sequential batch cycles

APB (cost)	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
M1		− 9204		− 3269			− 2725			4022
M2			2262	33	1865			1746		2790
M3			392	− 6332				105		− 902
M4			− 110				4033			4555
M5		− 78		5762	5934		3687	1758	− 1640	
M6			526	− 2332				336		151
M7				6490			1951			2250

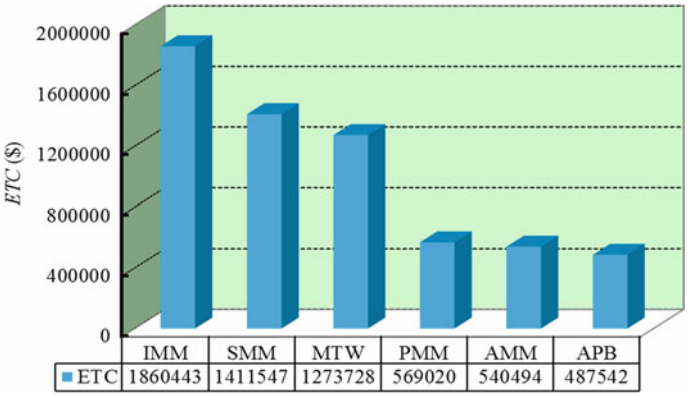


Fig. 8 Results comparison of opportunistic maintenance policies

6.5 *Effectiveness of RMTW Policy*

The RMTW policy is performed on a reconfigurable manufacturing CPS with changeable system structures shown in Fig. 4. In the first manufacturing stage (MS_1), $T_{W1} = 800$ is applied for the RMTW programming as an example, while $T_{W2} = 600$ and $T_{W3} = 1000$ are taken for MS_2 and MS_3 separately. Table 6 shows the RMTW scheduling results for reconfigured system structures. At each system-level PM execution point t_k, $\Theta(j, t_k) = 0$ means no PM action but this machine will be down according to the system structure; $\Theta(j, t_k) = 1$ indicates a PM action is combined to be performed; while $\Theta(j, t_k) = 2$ evinces no PM and this machine continues working. Newly added or removed machines are considered in each manufacturing stage.

From the results of reconfigurable manufacturing CPS, we can find that different CPS structures with various machine reliabilities and changeable system-level reconfigurations would lead to different ETC-saving rates. However, RMTW policy is exactly designed to redefine the time width of T_{Wh} for minimizing the ETC in each manufacturing stage. Therefore, this optimization mechanism ensures that RMTW policy can not only be rapidly adapt to new diverse system structures, but also achieve cost effectiveness for the whole-CPS maintenance scheduling. In Fig. 9, results indicate that the ETC-saving rate (28.105% comparing to IMM) achieved by RMTW scheduling is much higher than traditional opportunistic maintenance policies (IMM, SMM and static MTW). It can be concluded that proposed RMTW policy is a viable and effective policy to achieve rapid responsiveness and cost reduction for future reconfigurable manufacturing.

Table 6 RMTW results for reconfigurable manufacturing CPS

$\Theta(j, t_k)$	MS_1				MS_2					MS_3		
t_k	4587	6042	9033	10,251	12,563	16,012	18,437	19,147	19,978	21,955	24,188	27,538
M1	1	2	1	0	–	–	–	–	–	–	–	–
M2	1	0	0	1	0	1	2	0	0	1	0	1
M3	0	1	0	0	1	0	2	1	0	–	–	–
M4	0	2	1	0	2	0	1	2	0	2	0	1
M5	1	2	0	1	2	1	2	2	0	1	0	1
M6	–	–	–	–	2	1	2	2	1	2	1	1
M7	–	–	–	–	2	0	2	2	1	2	0	1
M8	–	–	–	–	–	–	–	–	–	2	0	1

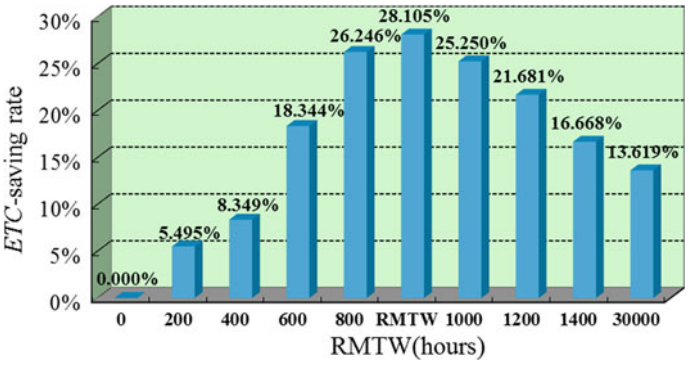


Fig. 9 ETC-saving rate comparison with various methods

7 Conclusions and Perspectives

In this chapter, we have presented systematic PHM methodologies for cyber-physical systems of three advanced manufacturing paradigms. With monitored and synchronized information from the cyber computational space, PHM methodologies integrating manufacturing characters in the physical factory floor can improve the health management. These developed prognosis algorithm, scheduling model and opportunistic maintenance policies achieve significant improvements in following aspects: (1) WFRGM algorithm provides real-time and accurate health predictions by incorporating updated information and influencing factors; (2) MAM method can output sequential PM intervals based on individual machine health for supporting the system-level opportunistic maintenance; (3) MTW policy schedules PM separations/combinations according to series–parallel structures for reducing maintenance cost and decision-making complexity; (4) APB policy achieves huge cost savings by utilizing set-up times to makes real-time PM optimizations and handle variable batch orders; (5) RMTW policy efficiently achieves rapid responsiveness and cost effectiveness for diverse open-ended reconfigurations and flexible system structures.

In sum, both cyber factors (information technologies) and physical factors (manufacturing paradigms) are essential for the health management of future CPS. Some industrial enterprises (i.e., port machinery manufacturers and automobile manufacturing companies) have already benefited from these novel PHM methodologies. Future work is needed to extending this hierarchical PHM framework to other burgeoning manufacturing paradigms, such as sustainable manufacturing, green production and cloud manufacturing.

References

1. Lee J, Kao HA, Yang S (2014) Service innovation and smart analytics for industry 4.0 and big data environment. Procedia CIRP 16:3–8
2. Xia T, Dong Y, Xiao L, Du S, Pan E, Xi L (2018) Recent advances in prognostics and health management for advanced manufacturing paradigms. Reliab Eng Syst Saf 178:255–268
3. Lin QL, Liu HC, Wang DJ, Liu L (2015) Integrating systematic layout planning with fuzzy constraint theory to design and optimize the facility layout for operating theatre in hospitals. J Intell Manuf 26(1):87–95
4. Jardim-Goncalves R, Grilo A, Popplewell K (2016) Novel strategies for global manufacturing systems interoperability. J Intell Manuf 27(1):1–9
5. Rafiee K, Feng Q, Coit DW (2014) Reliability modeling for dependent competing failure processes with changing degradation rate. IIE Trans 46(5):483–496
6. Benkedjouh T, Medjaher K, Zerhouni N, Rechak S (2015) Health assessment and life prediction of cutting tools based on support vector regression. J Intell Manuf 26(2):213–223
7. Arab A, Ismail N, Lee LS (2013) Maintenance scheduling incorporating dynamics of production system and real-time information from workstations. J Intell Manuf 24(4):695–705
8. Mirabi M, Ghomi SMT, Jolai F (2013) A two-stage hybrid flowshop scheduling problem in machine breakdown condition. J Intell Manuf 24(1):193–199
9. Zied H, Sofiene D, Nidhal R (2014) Joint optimisation of maintenance and production policies with subcontracting and product returns. J Intell Manuf 25(3):589–602
10. Liu X, Li J, Al-Khalifa KN, Hamouda AS, Coit DW, Elsayed EA (2013) Condition-based maintenance for continuously monitored degrading systems with multiple failure modes. IIE Trans 45(4):422–435
11. Sheikhalishahi M, Ebrahimipour V, Farahani MH (2014) An integrated GA-DEA algorithm for determining the most effective maintenance policy for ak-out-of-n problem. J Intell Manuf 25(6):1455–1462
12. Chouikhi H, Khatab A, Rezg N (2014) A condition-based maintenance policy for a production system under excessive environmental degradation. J Intell Manuf 25(4):727–737
13. Ebrahimipour V, Najjarbashi A, Sheikhalishahi M (2015) Multi-objective modeling for preventive maintenance scheduling in a multiple production line. J Intell Manuf 26(1):111–122
14. Xia T, Zhang K, Sun B, Fang X, Xi L (2021) Integrated remanufacturing and opportunistic maintenance decision-making for leased batch production lines. J Manuf Sci Eng 143(8)
15. Xia T, Sun B, Chen Z, Pan E, Wang H, Xi L (2021) Opportunistic maintenance policy integrating leasing profit and capacity balancing for serial-parallel leased systems. Reliab Eng Syst Saf 205:107233
16. Zhou X, Xi L, Lee J (2009) Opportunistic preventive maintenance scheduling for a multi-unit series system based on dynamic programming. Int J Prod Econ 118(2):361–366
17. Lee S, Gu X, Ni J (2013) Stochastic maintenance opportunity windows for unreliable two-machine one-buffer system. Expert Syst Appl 40(13):5385–5394
18. Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D (2014) Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. Mech Syst Signal Process 42(1–2):314–334
19. Lee J, Bagheri B, Kao HA (2015) A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manuf Lett 3:18–23
20. Pasqualetti F, Dörfler F, Bullo F (2013) Attack detection and identification in cyber-physical systems. IEEE Trans Autom Control 58(11):2715–2729
21. Monostori L et al (2016) Cyber-physical systems in manufacturing. Cirp Annals 65(2):621–641
22. Wang L, Törngren M, Onori M (2015) Current status and advancement of cyber-physical systems in manufacturing. J Manuf Syst 37:517–527
23. Gao R, Wang L, Teti R, Dornfeld D, Kumara S, Mori M, Helu M (2015) Cloud-enabled prognosis for manufacturing. CIRP Ann 64(2):749–772
24. Wang CH, Hsu LC (2008) Using genetic algorithms grey theory to forecast high technology industrial output. Appl Math Comput 195(1):256–263

25. Yu J, Xi L (2008) Intelligent monitoring and diagnosis of manufacturing process using an integrated approach of neural network ensemble and genetic algorithm. Int J Comput Appl Technol 33(2–3):109–119
26. Tian Z (2012) An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. J Intell Manuf 23(2):227–237
27. Akay D, Atak M (2007) Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. Energy 32(9):1670–1675
28. Xia T, Jin X, Xi L, Zhang Y, Ni J (2015) Operating load based real-time rolling grey forecasting for machine health prognosis in dynamic maintenance schedule. J Intell Manuf 26(2):269–280
29. Ju-Long D (1982) Control problems of grey systems. Syst Control Lett 1(5):288–294
30. Xia T, Ding Y, Dong Y, Chen Z, Zheng M, Pan E, Xi L (2022) Collaborative production and predic-tive maintenance scheduling for flexible flow shop with stochastic inter-ruptions and monitoring data. J Manuf Syst 65:640–652
31. Xia T, Fang X, Gebraeel N, Xi L, Pan E (2019) Online analytics framework of sensor-driven prognosis and opportunistic maintenance for mass customization. J Manuf Sci Eng 141(5)
32. Wang CH, Tsai SW (2014) Optimizing bi-objective imperfect preventive maintenance model for series-parallel system using established hybrid genetic algorithm. J Intell Manuf 25(3):603–616
33. Xia T, Xi L, Zhou X, Lee J (2012) Dynamic maintenance decision-making for series–parallel manufacturing system based on MAM–MTW methodology. Eur J Oper Res 221(1):231–240
34. Xia T, Dong Y, Pan E, Zheng M, Wang H, Xi L (2021) Fleet-level opportunistic maintenance for large-scale wind farms integrating real-time prognostic updating. Renew Energy 163:1444–1454
35. Xia T, Jin X, Xi L, Ni J (2015) Production-driven opportunistic maintenance for batch production based on MAM–APB scheduling. Eur J Oper Res 240(3):781–790
36. Dong Y, Xia T, Fang X, Zhang Z, Xi L (2019) Prognostic and health management for adaptive manufacturing systems with online sensors and flexible structures. Comput Ind Eng 133:57–68
37. Ni J, Jin X (2012) Decision support systems for effective maintenance operations. CIRP Ann 61(1):411–414
38. Xia T, Xi L, Pan E, Ni J (2017) Reconfiguration-oriented opportunistic maintenance policy for reconfigurable manufacturing systems. Reliab Eng Syst Saf 166:87–98

Degradation Modeling and Residual Life Prediction Based on Nonlinear Wiener Process

Bo Guo

Abstract Residual life estimation plays a significant role in scheduling maintenance activities for high-reliability products. In the literature, most of the existing studies dealt with this issue by considering only one-dimensional performance characteristic. However, it may be unreasonable since a product can have multiple performance characteristics. Generally, these performance characteristics are dependent due to the common influences from the environments. Moreover, the nonlinearity of the product's degradation process should also be taken into account. In this chapter, degradation models based on nonlinear Wiener process is presented to address the issue under univariate and multivariate situations. Based on the proposed method, a closed-form of the probability density function (PDF) of the product's residual life can be approximately obtained. Numerical examples concerning fatigue cracks demonstrate the validity of the proposed method.

Keywords Residual life estimation · Performance characteristic · Nonlinearity · Wiener process

1 Introduction

Due to the internal and external environments, the product's performance can degrade over time, leading to reduced reliability. The lower the reliability, the higher the failure probability of product. With respect to high-reliability products such as manned spaceships, satellites and aircraft engines, unexpected failures may lead to loss of human lives, negative impact on the environment, and financial loss. Therefore, the residual life (RL) of a product needs to be precisely predicted, and then served as an important basis for reliability improvement, failure prevention, and utilization maximization of the product through implementing optimal maintenance actions. Generally, a product's RL can be defined as the time length between the beginning and the end of its finite life. From the degradation point of view, a product can be viewed

B. Guo (✉)
College of Systems Engineering, National University of Defense Technology, Hunan, China
e-mail: boguo@nudt.edu.cn

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_18

as failure when its degradation value of the healthy state exceeds the corresponding failure threshold [1, 2]. In this regard, the degradation process and the product's RL should be appropriately described and predicted. Generally. degradation models can be developed based on exponential models [3], autoregressive models [4], stochastic process models [5], neural network models [6], and so on.

Considering the random effects induced by the operating environments, The stochastic process models are usually employed to model the degradation of a product [7]. Among them, the Wiener process is widely used in degradation analysis owing to its useful mathematical properties. For the linear degradation processes, the Wiener process with a linear drift has a wide application [5]. When it comes to nonlinear degradation processes, the non-linearity should be fully considered in degradation modeling. Otherwise, the dynamics during the degradation process cannot be properly captured, which can further lead to wrong results of the RL prediction. Therefore, we pay more attentions to the degradation modeling with a nonlinear Wiener process. Moreover, most of the existing studies on RL prediction assumed that the product's health state can be completely characterized by a performance characteristic (PC) [3–5, 7–9]. In fact, a product may have multiple performance characteristics (PCs), most importantly, the degradations of these PCs may be dependent due to the common effects from the environments [10–13]. For the sake of estimating the RL of a product with multiple PCs, it is necessary to construct multivariate degradation models to properly capture the degradation behaviors of the PCs.

2 Degradation Modeling and Residual Life Prediction with a Univariate Nonlinear Wiener Process

Suppose that one PC with nonlinear degradation behavior can characterize a product's health state. When the value of the PC exceeds the degradation threshold D, the target product is regarded as failed. Next, we will describe the degradation modeling principle and the probability density function (PDF) of the RL using a nonlinear Wiener process.

2.1 Degradation Modeling

Denote $X(t)$ as the cumulative degradation of a PC at time t, and it is written as:

$$M_1 : X(t) = a\Lambda(t;\boldsymbol{\theta}) + \sigma B(\tau(t;\boldsymbol{\gamma})) \tag{1}$$

where both $\Lambda(t;\boldsymbol{\theta})$ and $\tau(t;\boldsymbol{\gamma})$ are linear/nonlinear consecutively increasing functions in $t \in [0,\infty)$. $B(\cdot)$ denotes the Brownian motion (BM), and it is utilized to characterize the temporal uncertainty of the degradation process. $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and σ are

fixed effects that describe the population's common degradation characteristics. a is a random effect that captures the product-to-product variability. Here, a is supposed to be normally distributed with mean μ_0 and variance σ_0^2.

Owing to the measurement error or other types of factors, the degradation process can be polluted by noise. In this regard, the degradation measurement at time t can be calculated by adding a measurement error $\varepsilon(t)$ to $X(t)$, formulated as:

$$M_2 : Y(t) = X(t) + \varepsilon(t) \tag{2}$$

where $\varepsilon(t)$ obeys a normal distribution with mean 0 and variance ξ^2. Furthermore, the measurement errors at different time instances are assumed to be independent, that is,

$$\operatorname{cov}\left(\varepsilon\left(t_j\right), \varepsilon(t_l)\right) = \begin{cases} \xi^2, & j = l; \\ 0, & j \neq l. \end{cases}$$

Particularly, the proposed model in Eq. (2) encompasses the following special cases:

(1) If $\xi \to 0$, M_2 reduces to M_1;
(2) Given $\Lambda(t; \boldsymbol{\theta}) = \tau(t; \boldsymbol{\gamma})$, the generalized Wiener process becomes a regular Wiener process with a time-scale transformation [10];
(3) If $\Lambda(t; \boldsymbol{\theta}) = \tau(t; \boldsymbol{\gamma})$ and $\sigma_0 \to 0$, M_2 reduces to the model proposed in [14];
(4) If $\Lambda(t; \boldsymbol{\theta}) = \tau(t; \boldsymbol{\gamma}) = t$, M_2 becomes the model studied in [15];
(5) By specifying $\Lambda(t; \boldsymbol{\theta})$ and $\tau(t; \boldsymbol{\gamma})$ in different forms, M_2 can be extended to cover different degradation models.

Let $t_0 = 0, \Delta Y_1 = Y(t_1)$, $\Delta Y_j = Y(t_j) - Y(t_{j-1})$, $\Delta\Lambda(t_j; \boldsymbol{\theta}) = \Lambda(t_j; \boldsymbol{\theta}) - \Lambda(t_1; \boldsymbol{\theta})$, and $\Delta\tau(t_j; \boldsymbol{\gamma}) = \tau(t_j; \boldsymbol{\gamma}) - \tau(t_{j-1}; \boldsymbol{\gamma})$. Given a, we have

$$\operatorname{cov}\left(\Delta Y_j, \Delta Y_l | a\right) = \begin{cases} \sigma^2 \Delta\tau\left(t_j; \boldsymbol{\gamma}\right) + \xi^2, & j = l = 1, \\ \sigma^2 \Delta\tau\left(t_j; \boldsymbol{\gamma}\right) + 2\xi^2, & j = l > 1, \\ -\xi^2, & l = j + 1 \text{ or } l = j - 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Furthermore, the increment of the degradation measurement ΔY_j, given a, is normally distributed, that is,

$$\Delta Y_j | a \sim \begin{cases} N\left(a\Delta\Lambda\left(t_j; \boldsymbol{\theta}\right), \sigma^2 \Delta\tau\left(t_j; \boldsymbol{\gamma}\right) + \xi^2\right), & j = 1, \\ N\left(a\Delta\Lambda\left(t_j; \boldsymbol{\theta}\right), \sigma^2 \Delta\tau\left(t_j; \boldsymbol{\gamma}\right) + 2\xi^2\right), & j > 1. \end{cases} \tag{4}$$

Let $\Delta \boldsymbol{Y} = (\Delta Y_1, \Delta Y_2, \ldots, \Delta Y_n)'$, $\Delta \boldsymbol{\Lambda} = (\Delta\Lambda(t_1; \boldsymbol{\theta}), \Delta\Lambda(t_2; \boldsymbol{\theta}), \ldots, \Delta\Lambda(t_n; \boldsymbol{\theta}))'$. According to Eqs. (3) and (4), we have

$$\Delta \boldsymbol{Y} | a \sim N(a\Delta \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \tag{5}$$

with the joint PDF as:

$$f_{\Delta \boldsymbol{Y}}(\Delta \boldsymbol{Y}|a) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{(\Delta \boldsymbol{Y} - a\Delta \boldsymbol{\Lambda})' \boldsymbol{\Sigma}^{-1}(\Delta \boldsymbol{Y} - a\Delta \boldsymbol{\Lambda})}{2}\right)$$

where $\boldsymbol{\Sigma}$ is the conditional variance–covariance matrix of $\Delta \boldsymbol{Y}$ with the (j, l)th element as $\text{cov}(\Delta Y_j, \Delta Y_l|a)$.

2.2 Degradation Model Updating

To accurately predict a product's RL, the future degradation behavior is expected to be properly captured. Hence, the degradation model is updated using the historical degradation data at each time point, making it possible to characterize the real degradation process as close as possible. Subsequently, the RL can be timely estimated based on the updated degradation model and historical data.

Assume that a product is not failed at current time t_h and denote the true cumulative degradation value as $X(t_h)$. One can represent the degradation process of a product as:

$$X(t_h + l) = X(t_h) + a\Delta\Lambda(t_h + l; \boldsymbol{\theta}) + \sigma B(\Delta\tau(t_h + l; \boldsymbol{\gamma})) \tag{6}$$

where $\Delta\Lambda(t_h + l; \boldsymbol{\theta}) = \Lambda(t_h + l; \boldsymbol{\theta}) - \Lambda(t_h; \boldsymbol{\theta})$ and $\Delta\tau(t_h + l; \boldsymbol{\gamma}) = \tau(t_h + l; \boldsymbol{\gamma}) - \tau(t_h; \boldsymbol{\gamma})$.

Given $X(t_h)$, the dynamics of $X(t_h + l)$ can be captured. However, to reduce the uncertainty in estimation, the extrapolation of the future degradation should be timely related to the degradation history, rather than the current degradation value [5]. Thus, we reconstruct Eq. (6) as:

$$\begin{cases} a_j = a_{j-1} + \eta, \\ X\left(t_j\right) = X\left(t_{j-1}\right) + a_{j-1}\Delta\Lambda\left(t_j; \boldsymbol{\theta}\right) + \varsigma_j, \end{cases} \tag{7}$$

where $\eta \sim N(0, Q)$ and $\varsigma_j \sim N(0, \sigma^2\Delta\tau(t_j; \boldsymbol{\gamma}))$. Based on degradation history $\boldsymbol{X}_{1:h} = (X(t_1), X(t_2), \ldots, X((t_h)))$, a_h could be estimated by Eq. (7). The dynamics of $X(t_h + l)$ can then be denoted as:

$$X(t_h + l) = X(t_h) + a_h\Delta\Lambda(t_h + l; \boldsymbol{\theta}) + \sigma B(\Delta\tau(t_h + l; \boldsymbol{\gamma}))$$

For the sake of convenience, the entire process is referred as degradation model updating.

However, due to measurement errors, only the history of the degradation measurements $\boldsymbol{Y}_{1:h} = (Y(t_1), Y(t_2), \ldots, Y((t_h)))$ is known whereas $\boldsymbol{X}_{1:h}$ remains unknown. Therefore, the degradation model cannot be directly updated by Eq. (7). Thus, we rewrite Eq. (7) as:

$$\begin{cases} a_j = a_{j-1} + \eta, \\ X\left(t_j\right) = X\left(t_{j-1}\right) + a_{j-1}\Delta\Lambda\left(t_j; \boldsymbol{\theta}\right) + \varsigma_j, \\ Y\left(t_j\right) = X\left(t_j\right) + \varepsilon_j, \end{cases} \tag{8}$$

where $\varepsilon_j \sim N(0, \xi^2)$. Now, both a_j and $X(t_j)$ can be seen as hidden states.

To estimate the hidden states a_h and $X(t_h)$, we rewrite Eq. (8) as:

$$\begin{cases} \tilde{\boldsymbol{X}}_j = \boldsymbol{F}_{j,j-1}\tilde{\boldsymbol{X}}_{j-1} + \boldsymbol{v}\left(t_j\right), \\ Y\left(t_j\right) = \boldsymbol{H}\tilde{\boldsymbol{X}}_j + \varepsilon_j, \end{cases} \tag{9}$$

where $\tilde{\boldsymbol{X}}_j = (a_j, X(t_j))'$, $\boldsymbol{v}(t_j) = (\eta, \varsigma_j)'$, $\boldsymbol{H} = (0, 1)$, and $\boldsymbol{F}_{j,j-1} = \begin{bmatrix} 1 & 0 \\ \Delta\Lambda(t_j; \boldsymbol{\theta}) & 1 \end{bmatrix}$. The covariance matrix of $\boldsymbol{v}\left(t_j\right)$ is presented as:

$$\mathbf{M}_j = \begin{bmatrix} Q & 0 \\ 0 & \sigma^2\Delta\tau\left(t_j; \boldsymbol{\gamma}\right) \end{bmatrix}$$

Given Eq. (9), the hidden states can be estimated using a Kalman filter The following steps describe the main estimation steps.

Step 1: Initialize the hidden states

$$\hat{\tilde{\boldsymbol{X}}}_0 = (\mu_0, 0)', P_{0|0} = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & 0 \end{bmatrix}, j = 0$$

Step 2: Set $j = j + 1$, calculate the one-step prediction of $\tilde{\boldsymbol{X}}_j$ as:

$$\hat{\tilde{\boldsymbol{X}}}_{j+1|j} = \boldsymbol{F}_{j+1,j}\hat{\tilde{\boldsymbol{X}}}_j \tag{10}$$

$$\boldsymbol{P}_{j+1|j} = \boldsymbol{F}_{j+1,j}\boldsymbol{P}_{j|j}\boldsymbol{F}'_{j+1,j} + \mathbf{M}_j \tag{11}$$

Step 3: Estimate the hidden state as:

$$\hat{\tilde{\boldsymbol{X}}}_{j+1|j+1} = \hat{\tilde{\boldsymbol{X}}}_{j+1|j} + K_{j+1}\left(Y_{j+1} - \hat{Y}_{j+1|j}\right) \tag{12}$$

where $\hat{Y}_{j+1|j} = \boldsymbol{H}\hat{\tilde{\boldsymbol{X}}}_{j+1|j}$, $K_{j+1} = \boldsymbol{P}_{j+1|j}\boldsymbol{H}'(H\boldsymbol{P}_{j+1|j}H' + \xi^2)^{-1}$;

Step 4: Update the variance of the hidden state by:

$$\boldsymbol{P}_{j+1|j+1} = \left(I - K_{j+1}\boldsymbol{H}\right)\boldsymbol{P}_{j+1|j} \tag{13}$$

where $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$;

Step 5: Repeat step (2) to step (4) until $j = h$.

Ultimately, a_h and $X(t_h)$ can be estimated. Their uncertainty can be respectively reflected as:

$$a_h \sim N\left(\hat{a}_h, \boldsymbol{P}^{1,1}_{h|h}\right), X(t_h) \sim N\left(\widehat{X(t_h)}, \boldsymbol{P}^{2,2}_{h|h}\right) \tag{14}$$

where $(\hat{a}_h, \widehat{X(t_h)})' = \hat{\tilde{\boldsymbol{X}}}_h$ and $\boldsymbol{P}^{l,l}_{h|h}$ is the (l, l)th element of $\boldsymbol{P}_{h|h}$. When new degradation measurement information becomes available, we can again estimate the hidden states by the above procedures.

2.3 Residual Life Prediction

Given the updated degradation model, a product's RL at the current time point is represented as:

$$L_h = \inf\{l : X(t + l) \geq D\} \tag{15}$$

Let $Z(l) = X(t_h + l) - X(t_h)$, i.e.,

$$Z(l) = a_h \Delta\Lambda(t_h + l; \boldsymbol{\theta}) + \sigma B(\Delta\tau(t_h + l; \boldsymbol{\gamma}))$$

Clearly, L_h is the first passage time that $Z(l)$ exceeds $D_h = D - X(t_h)$. Since the nonlinearity may exist in either $\Lambda(t; \boldsymbol{\theta})$ or $\tau(t; \boldsymbol{\gamma})$, it is difficult to obtain the PDF of L_h in a closed-form. Given a_h and $X(t_h)$, the PDF of L_h can be approximated as [7]:

$$f_{L_h}(l|a_h, X(t_h)) \cong \frac{1}{A_{L_h}} g_{L_h}(l|a_h, X(t_h)) \tag{16}$$

where

$$g_{L_h}(l|a_h, X(t_h)) = \frac{1}{\sqrt{2\pi\,\Delta\tau(t_h + l; \boldsymbol{\gamma})}}$$

$$\left(\frac{D_{Y_h}(\Delta\tau(t_h+l;\boldsymbol{\gamma}))}{\Delta\tau(t_h+l;\boldsymbol{\gamma})}+\frac{a_h\eta_h'(\Delta\tau(t_h+l;\boldsymbol{\gamma});\boldsymbol{\theta})}{\sigma}\right)\times\exp\left(-\frac{\left(D_{Y_h}(\Delta\tau(t_h+l;\boldsymbol{\gamma}))\right)^2}{2\Delta\tau(t_h+l;\boldsymbol{\gamma})}\right)\frac{d\Delta\tau(t_h+l;\boldsymbol{\gamma})}{dl},$$

$A_{L_h}=\int_0^\infty g_{L_h}(l|a_h)dl$, $D_{Y_h}(s_h)=\frac{1}{\sigma}(D_h-a_h\eta_h(s_h;\boldsymbol{\theta}))$, $\eta_h(s_h;\boldsymbol{\theta})=\Delta\Lambda(t_h+\nu(s_h;\boldsymbol{\theta});\boldsymbol{\theta})$, $\nu(s_h;\boldsymbol{\gamma})$ is the inverse of $\Delta\tau(t_h+l;\boldsymbol{\gamma})$ by setting $s_h=\Delta\tau(t_h+l;\boldsymbol{\gamma})$, and $a_h\eta_h'(s_h;\boldsymbol{\theta})=\frac{d\eta_h(s_h;\boldsymbol{\theta})}{ds_h}$.

Then, the unconditional PDF of L_h can be formulated by using the law of total probability:

$$f_{L_h}(l)\cong\iint_{\Omega}f_{L_h}(l|a_h,X(t_h))da_hdX(t_h)\tag{17}$$

However, the closed-form of $f_{L_h}(l)$ is hard to be derived owing to A_{L_h}. Therefore, it is necessary to obtain a new approximation.

Suppose that $Z\sim N(\mu_Z,\sigma_Z^2)$, we the have

$$E_Z\left(\exp\left(-\frac{(K_1-K_2Z)^2}{2K_3}\right)\right)=\sqrt{\frac{K_3}{(K_2\sigma_Z)^2+K_3}}\exp\left(-\frac{(K_1-K_2\mu_Z)^2}{2\left((K_2\sigma_Z)^2+K_3\right)}\right)\tag{18}$$

and

$$\begin{aligned}&E_Z\left(Z\exp\left(-\frac{(K_1-K_2Z)^2}{2K_3}\right)\right)\\&=\frac{(\sigma_Z)^2K_2K_1+\mu_ZK_3}{(K_2\sigma_Z)^2+K_3}\sqrt{\frac{K_3}{(K_2\sigma_Z)^2+K_3}}\exp\left(-\frac{(K_1-K_2\mu_Z)^2}{2\left((K_2\sigma_Z)^2+K_3\right)}\right).\end{aligned}\tag{19}$$

Here, the law of total probability is firstly applied on $g_{L_h}(l|a_h,X(t_h))$ with respect to a_h:

$$\begin{aligned}g_{L_h}(l|X(t_h))=&\frac{1}{\Delta\tau(t_h+l;\boldsymbol{\gamma})\sqrt{2\pi\left(B^2\boldsymbol{P}_{h|h}^{1,1}+C\right)}}\frac{d\Delta\tau(t_h+l;\boldsymbol{\gamma})}{dl}\\&\times\left(D_h-A\frac{B\boldsymbol{P}_{h|h}^{1,1}D_h+\hat{a}_hC}{B^2\boldsymbol{P}_{h|h}^{1,1}+C}\right)\exp\left(-\frac{\left(D_h-\hat{a}_hB\right)^2}{2\left(B^2\boldsymbol{P}_{h|h}^{1,1}+C\right)}\right),\end{aligned}\tag{20}$$

where $A=\eta_h(\Delta\tau(t_h+l;\boldsymbol{\gamma});\boldsymbol{\theta})-\eta_h'(\Delta\tau(t_h+l;\boldsymbol{\gamma});\boldsymbol{\theta})\Delta\tau(t_h+l;\boldsymbol{\gamma}),B=\eta_h(\Delta\tau(t_h+l;\boldsymbol{\gamma});\boldsymbol{\theta}),C=\sigma^2\Delta\tau(t_h+l;\boldsymbol{\gamma})$. Since $D_h=D-X(t_h)$ and $X(t_h)\sim(\widehat{X(t_h)},\boldsymbol{P}_{h|h}^{2,2})$,

$D_h \sim N(\overline{D}_h, \boldsymbol{P}^{2,2}_{h|h})$. By integrating D_h out of $g_{L_h}(l|X(t_h))$, we get

$$g_{L_h}(l) = \frac{1}{\Delta\tau(t_h+l;\boldsymbol{\gamma})\sqrt{2\pi\left(B^2\boldsymbol{P}^{1,1}_{h|h}+C\right)}}\frac{d\Delta\tau(t_h+l;\boldsymbol{\gamma})}{dl} \times\left(\left(1-\frac{AB\boldsymbol{P}^{1,1}_{h|h}}{B^2\boldsymbol{P}^{1,1}_{h|h}+C}\right)\frac{\boldsymbol{P}^{2,2}_{h|h}\hat{a}_h B+\overline{D}_h\left(B^2\boldsymbol{P}^{1,1}_{h|h}+C\right)}{\boldsymbol{P}^{2,2}_{h|h}+B^2\boldsymbol{P}^{1,1}_{h|h}+C}-\frac{A\hat{a}_h C}{B^2\boldsymbol{P}^{1,1}_{h|h}+C}\right) \times\sqrt{\frac{B^2\boldsymbol{P}^{1,1}_{h|h}+C}{\boldsymbol{P}^{2,2}_{h|h}+B^2\boldsymbol{P}^{1,1}_{h|h}+C}}\exp\left(-\frac{\left(\hat{a}_h B-\overline{D}_h\right)^2}{2\left(\boldsymbol{P}^{2,2}_{h|h}+B^2\boldsymbol{P}^{1,1}_{h|h}+C\right)}\right). \tag{21}$$

Finally, we approximate $f_{L_h}(l)$ as:

$$f_{L_h}(l) \cong \frac{1}{A'_{L_h}}g_{L_h}(l) \tag{22}$$

where $A'_{L_h} = \int_0^\infty g_{L_h}(l)dl$. Based on Eq. (22), $f_{L_h}(l)$ can be approximately expressed in a closed-form. For example, given $\Lambda(t;\boldsymbol{\theta}) = t^b$ and $\tau(t;\boldsymbol{\gamma}) = t^\gamma$, $f_{L_h}(l)$ is written as:

$$f_{L_h}(l) \approx \frac{1}{A'_{L_h}}\frac{\gamma(t_h+l)^{\gamma-1}}{\Delta\tau(t_h+l;\boldsymbol{\gamma})\sqrt{2\pi B_0}}\sqrt{\frac{B_0}{B_0+\boldsymbol{P}^{2,2}_{h|h}}}\exp\left(-\frac{\left(\bar{D}_h-\hat{a}_h\Delta\Lambda(t_h+l;\boldsymbol{\theta})\right)^2}{2\left(B_0+\boldsymbol{P}^{2,2}_{h|h}\right)}\right) \times\left(\left(1-\frac{A\Delta\Lambda(t_h+l;\boldsymbol{\theta})\boldsymbol{P}^{1,1}_{h|h}}{B_0}\right)\frac{\boldsymbol{P}^{2,2}_{h|h}\hat{a}_h\Delta\Lambda(t_h+l;\boldsymbol{\theta})+\bar{D}_h B_0}{\boldsymbol{P}^{2,2}_{h|h}+B_0} -\frac{A\hat{a}_h\sigma^2\Delta\tau(t_h+l;\boldsymbol{\gamma})}{B_0}\right), \tag{23}$$

where $A = \Delta\Lambda(t_h+l;\boldsymbol{\theta}) - \frac{b(t_h+l)^{b-\gamma}\Delta\tau(t_h+l;\boldsymbol{\gamma})}{\gamma}$ and $\mathrm{B}_0 = \Delta\Lambda(t_h+l;\boldsymbol{\theta})^2\boldsymbol{P}^{1,1}_{h|h} + \sigma^2\Delta\tau(t_h+l;\boldsymbol{\gamma})$.

Until now, given $\boldsymbol{Y}_{1:h}$, we can estimate the PDF of L_h by Eq. (22).

2.4 Parameters Estimation

Before estimating a product's RL, some parameters, denoted as $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \mu_0, \sigma_0, \sigma, \xi, Q)$, should be determined. To this end, we divide $\boldsymbol{\Theta}$ into two parts, including $\boldsymbol{\Theta}_1 = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \mu_0, \sigma_0, \xi, \sigma)$ and Q. Specifically, $\boldsymbol{\Theta}_1$ characterizes the population-based degradation characteristics and Q describes the product's dynamic

characteristics. Therefore, to estimate $\boldsymbol{\Theta}_1$ and Q, the population-based degradation information and the historical degradation information are employed, respectively.

Suppose that the population-based degradation information is composed of m products' degradation measurements, where the degradation value of the ith product is tested at n_i times. Denote the degradation measurement information of the ith product in the jth test as $Y_i(t_{i,j})$ $(i = 1, 2, \ldots, m, j = 1, 2, \ldots, n_i)$. Let $\Delta \boldsymbol{Y}_i = (\Delta Y_{i,1}, \Delta Y_{i,2}, \ldots, \Delta Y_{i,n_i})'$ and $\Delta \boldsymbol{\Lambda}_i = (\Delta \Lambda(t_{i,1}; \boldsymbol{\theta}), \Delta \Lambda(t_{i,2}; \boldsymbol{\theta}), \ldots, \Delta \Lambda((t_{i,n_i}; \boldsymbol{\theta})))'$, where $\Delta Y_{i,j} = Y_i(t_{i,j}) - Y_i(t_{i,j-1})$, $\Delta \Lambda(t_{i,j}; \boldsymbol{\theta}) = \Lambda(t_{i,j}; \boldsymbol{\theta}) - \Lambda(t_{i,j-1}; \boldsymbol{\theta})$. According to Eq. (5), we have:

$$\Delta \boldsymbol{Y}_i | a_{(i)} \sim N\left(a_{(i)} \boldsymbol{\Lambda}_i, \boldsymbol{\Sigma}_i\right)$$

where $a_{(i)}$ is a realization of a with respect to the ith product, and $\boldsymbol{\Sigma}_i$ is the covariance matrix of $\Delta \boldsymbol{Y}_i$ conditional on $a_{(i)}$. The (j, l)th of $\boldsymbol{\Sigma}_i$ is given as:

$$\boldsymbol{\Sigma}_{i,j,l} = \begin{cases} \sigma^2 \Delta \tau\left(t_{i,j}; \boldsymbol{\gamma}\right) + \xi^2, & j = l = 1, \\ \sigma^2 \Delta \tau\left(t_{i,j}; \boldsymbol{\gamma}\right) + 2\xi^2, & j = l > 1, \\ -\xi^2, & l = j + 1 \text{ or } l = j - 1, \\ 0, & \text{otherwise.} \end{cases}$$

where $\Delta \tau(t_{i,j}; \boldsymbol{\gamma}) = \tau(t_{i,j}; \boldsymbol{\gamma}) - \tau(t_{i,j-1}; \boldsymbol{\gamma})$. Then, conditional on $a_{(i)}$, the moment generating function of $\Delta \boldsymbol{Y}_i$ is given as:

$$M_{\Delta \boldsymbol{Y}_i | a_{(i)}}(\boldsymbol{w}_i) = E\left(\exp\left(\boldsymbol{w}'_i \Delta \boldsymbol{Y}_i\right)\right) = \exp\left(a_{(i)} \boldsymbol{w}'_i \Delta \boldsymbol{\Lambda}_i\right) \exp\left(\frac{\boldsymbol{w}'_i \boldsymbol{\Sigma}_i \boldsymbol{w}_i}{2}\right) \tag{24}$$

Since $a_{(i)} \sim N(\mu_0, \sigma_0^2)$, the unconditional moment generating function of $\Delta \boldsymbol{Y}_i$ can be obtained by using the law of total probability, and it is formulated as:

$$M_{\Delta \boldsymbol{Y}_i}(\boldsymbol{w}_i) = \exp\left(\mu_0 \boldsymbol{w}'_i \Delta \boldsymbol{\Lambda}_i\right) \exp\left(\frac{\boldsymbol{w}'_i \left(\boldsymbol{\Sigma}_i + \sigma_0^2 \Delta \boldsymbol{\Lambda}_i \Delta \boldsymbol{\Lambda}'_i\right) \boldsymbol{w}_i}{2}\right) \tag{25}$$

Thus,

$$\Delta \boldsymbol{Y}_i \sim N\left(\mu_0 \Delta \boldsymbol{\Lambda}_i, \boldsymbol{\Sigma}_i + \sigma_0^2 \Delta \boldsymbol{\Lambda}_i \Delta \boldsymbol{\Lambda}'_i\right) \tag{26}$$

Let $\Delta \boldsymbol{Y} = (\Delta \boldsymbol{Y}'_1, \Delta \boldsymbol{Y}'_2, \ldots, \Delta \boldsymbol{Y}'_m)'$, the log-likelihood function of $\boldsymbol{\Theta}_1$ can then be calculated by:

$$\ell(\boldsymbol{\Theta}_1 | \Delta \boldsymbol{Y}) = -\sum_{i=1}^{m} \frac{n_i}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{m} \log\left(\left|\boldsymbol{\Sigma}_i + \sigma_0^2 \Delta \boldsymbol{\Lambda}_i \Delta \boldsymbol{\Lambda}'_i\right|\right)$$

$$-\sum_{i=1}^{m}\frac{(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)'\left(\boldsymbol{\Sigma}_i+\sigma_0^2\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)}{2}. \tag{27}$$

Set $\tilde{\sigma}^2=\frac{\sigma^2}{\sigma_0^2}$, $\tilde{\xi}^2=\frac{\xi^2}{\sigma_0^2}$, and $\tilde{\boldsymbol{\Sigma}}_i=\frac{\boldsymbol{\Sigma}_i}{\sigma_0^2}$, the log-likelihood function can be reconstructed by:

$$\begin{aligned}\ell\left(\tilde{\boldsymbol{\Theta}}_1|\Delta\boldsymbol{Y}\right)=&-\sum_{i=1}^{m}\frac{n_i}{2}\log(2\pi)-\log\left(\sigma_0^2\right)\frac{\sum_{i=1}^{m}n_i}{2}-\frac{1}{2}\sum_{i=1}^{m}\log\left(\left|\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right|\right)\\&-\sum_{i=1}^{m}\frac{(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)'\left(\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)}{2\sigma_0^2},\end{aligned}$$

where $\tilde{\boldsymbol{\Theta}}_1=(\boldsymbol{\theta},\boldsymbol{\gamma},\mu_0,\sigma_0,\tilde{\xi},\tilde{\sigma})$. Upon differentiating $\ell\left(\tilde{\boldsymbol{\Theta}}_1|\Delta\boldsymbol{Y}\right)$ with respect to μ_0 and σ_0^2, respectively, we obtain:

$$\frac{\partial\ell\left(\tilde{\boldsymbol{\Theta}}_1|\Delta\boldsymbol{Y}\right)}{\partial\mu_0}=\sum_{i=1}^{m}\frac{\Delta\boldsymbol{\Lambda}'_i\left(\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)}{\sigma_0^2} \tag{28}$$

$$\begin{aligned}\frac{\partial\ell\left(\tilde{\boldsymbol{\Theta}}_1|\Delta\boldsymbol{Y}\right)}{\partial\sigma_0^2}=&-\frac{\sum_{i=1}^{m}n_i}{2\sigma_0^2}+\sum_{i=1}^{m}\\&\frac{(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)'\left(\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}(\Delta \boldsymbol{Y}_i-\mu_0\Delta\boldsymbol{\Lambda}_i)}{2\sigma_0^4}\end{aligned} \tag{29}$$

For specific values of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\tilde{\xi}^2$, and $\tilde{\sigma}^2$, the maximum likelihood estimate (MLE) values of μ_0 and σ_0^2 can be obtained by setting Eqs. (28) and (29) to 0, respectively, i.e.,

$$\hat{\mu}_0=\frac{\sum_{i=1}^{m}\Delta\boldsymbol{\Lambda}'_i\left(\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}\Delta \boldsymbol{Y}_i}{\sum_{i=1}^{m}\Delta\boldsymbol{\Lambda}'_i\left(\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}\Delta\boldsymbol{\Lambda}_i} \tag{30}$$

$$\hat{\sigma}_0^2=\frac{\sum_{i=1}^{m}\left(\Delta \boldsymbol{Y}_i-\hat{\mu}_0\Delta\boldsymbol{\Lambda}_i\right)'\left(\tilde{\boldsymbol{\Sigma}}_i+\Delta\boldsymbol{\Lambda}_i\Delta\boldsymbol{\Lambda}'_i\right)^{-1}\left(\Delta \boldsymbol{Y}_i-\hat{\mu}_0\Delta\boldsymbol{\Lambda}_i\right)}{\sum_{i=1}^{m}n_i} \tag{31}$$

By substituting $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ into $\ell(\tilde{\boldsymbol{\Theta}}_1|\Delta \boldsymbol{Y})$, the profile likelihood function of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\tilde{\xi}^2$, and $\tilde{\sigma}^2$ can be yielded, it is written as:

$$\ell\left(\boldsymbol{\theta}, \boldsymbol{\gamma}, \tilde{\xi}, \tilde{\sigma}|\Delta \boldsymbol{Y}\right) = -\sum_{i=1}^{m} \frac{n_i}{2} \log(2\pi) - \frac{\sum_{i=1}^{m} n_i}{2} - \log\left(\hat{\sigma}_0^2\right) \frac{\sum_{i=1}^{m} n_i}{2} - \frac{1}{2} \sum_{i=1}^{m} \log\left(\left|\tilde{\boldsymbol{\Sigma}}_i + \Delta \boldsymbol{\Lambda}_i \Delta \boldsymbol{\Lambda}'_i\right|\right) \tag{32}$$

Using a multiple-dimensional search approach, the optimal estimates of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\tilde{\xi}^2$, and $\tilde{\sigma}^2$ can be estimated by maximizing $\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}, \tilde{\xi}, \tilde{\sigma}|\Delta \boldsymbol{Y})$. Based on these estimates, the values of $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ can be obtained according to Eqs. (30) and (31), respectively. In addition, the confidence intervals and standard errors for the MLE results can also be obtained by applying the parametric bootstrap method [16].

Let $\boldsymbol{a}_{0:h-1} = (a_0, a_1, a_2, \ldots, a_{h-1})$ and denote the product's historical degradation measurement as $\boldsymbol{Y}_{1:h}$. Thus, the likelihood function of $\boldsymbol{a}_{0:h-1}$ and Q are represented by:

$$L(\boldsymbol{Y}_{1:h}|Q, \boldsymbol{a}_{0:h-1}) = \prod_{j=1}^{h} f_{\Delta Y_j}\left(\Delta Y_j|a_{j-1}\right) \tag{33}$$

where $f_{\Delta Y_j}(\Delta Y_j|a_{j-1})$ denotes the conditional PDF of ΔY_j obtained from Eq. (5). Owing to the existence of $\boldsymbol{a}_{0:h-1}$, it is hard to directly estimate Q. Hence, we solve this problem using a Bayesian approach.

Suppose that the prior PDF of Q is $\pi(Q)$. Then, the joint prior PDF of $\boldsymbol{a}_{0:h-1}$ and Q is given by:

$$\pi(Q, \boldsymbol{a}_{0:h-1}) = \prod_{j=1}^{h-1} f_{a_j}\left(a_j|a_{j-1}, Q\right) f_{a_0}(a_0)\pi(Q) \tag{34}$$

where

$$f_{a_j}\left(a_j|a_{j-1}, Q\right) = \frac{1}{\sqrt{2\pi Q}} \exp\left(-\frac{\left(a_j - a_{j-1}\right)^2}{2Q}\right)$$

$$f_{a_0}(a_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(a_0 - \mu_0)^2}{2(\sigma_0)^2}\right)$$

According to the Bayesian formula, the joint posterior PDF of $\boldsymbol{a}_{0:h-1}$ and Q are represented by:

$$\pi(Q, \boldsymbol{a}_{0:h-1}|\boldsymbol{Y}_{1:h}) \propto L(\boldsymbol{Y}_{1:h}|Q, \boldsymbol{a}_{0:h-1})\pi(Q, \boldsymbol{a}_{0:h-1}) \tag{35}$$

With a non-informative prior for Q, the Markov chain Monte Carlo (MCMC) method can be utilized to generate multiple samples with respect to Q from $\pi(Q, \boldsymbol{a}_{0:h-1}|\boldsymbol{Y}_{1:h})$, and the mean of these samples is treated as the estimate of Q.

2.5 Degradation Modeling and Residual Life Prediction with a Multivariate Nonlinear Wiener Process

In this section, the degradation behavior of a product is supposed to be characterized by multiple PCs and the degradation behavior of each PC is nonlinear. Moreover, these PCs are dependent with each other due to the influence from the environments. The target product is viewed as failure when any of these PCs firstly exceeds the corresponding degradation threshold.

2.6 Degradation Modeling

A Wiener process with a nonlinear drift is assumed to characterize the degradation process of each PC. The cumulative degradation value of the kth PC at time point t can be represented as:

$$X^{(k)}(t) = \beta^{(k)}\Lambda^{(k)}\left(t; \boldsymbol{\gamma}^{(k)}\right) + \sigma^{(k)}B^{(k)}(t) \tag{36}$$

The dependent relationship between the PCs is represented as:

$$\boldsymbol{X}(t)|\boldsymbol{\beta} \sim N(\boldsymbol{A}(t)\boldsymbol{\beta}, t\boldsymbol{\Sigma}) \tag{37}$$

where $\boldsymbol{X}(t) = (X^{(1)}(t), X^{(2)}(t), \ldots, X^{(p)}(t))'$, $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(p)})'$, $\boldsymbol{A}(t) = (A_{k,l}(t))_{p\times p}$, $\boldsymbol{\Sigma} = (\Sigma_{k,l})_{p\times p}$, $A_{k,l}(t) = \begin{cases} \Lambda^{(k)}(t; \boldsymbol{\gamma}^{(k)}), & k = l \\ 0, & otherwise \end{cases}$, and $\Sigma_{k,l} = \begin{cases} (\sigma^{(k)})^2, & k = l \\ \rho_{k,l}\sigma^{(k)}\sigma^{(l)}, & otherwise \end{cases}$. In order to characterize the degradation variabilities between different product, we assume that $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$.

Denote $\boldsymbol{X}_i(t_{i,j}) = (X_i^{(1)}(t_{i,j}), X_i^{(2)}(t_{i,j}), \ldots, X_i^{(p)}(t_{i,j}))'$ as the degradation measurement of product i at time point $t_{i,j}$ ($i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$). Set $\Delta t_{i,j} = t_{i,j} - t_{i,j-1}$, $\Delta X_i^{(k)}(t_{i,j}) = X_i^{(k)}(t_{i,j}) - X_i^{(k)}(t_{i,j-1})$, $\Delta\boldsymbol{A}(t_{i,j}) = \boldsymbol{A}(t_{i,j}) - \boldsymbol{A}(t_{i,j-1})$, and $\Delta\boldsymbol{X}_i(t_{i,j}) = (\Delta X_i^{(1)}(t_{i,j}), \Delta X_i^{(2)}(t_{i,j}), \ldots, \Delta X_i^{(p)}(t_{i,j}))'$. Thus, we have:

$$\Delta\boldsymbol{X}_i\left(t_{i,j}\right)|\boldsymbol{\beta}_i \sim N\left(\Delta\boldsymbol{A}\left(t_{i,j}\right)\boldsymbol{\beta}_i, \Delta t_{i,j}\boldsymbol{\Sigma}\right)$$

With $\boldsymbol{X}_i = ((\boldsymbol{X}_i(t_{i,1}))', (\boldsymbol{X}_i(t_{i,2}))', \ldots, (\boldsymbol{X}_i(t_{i,n_i}))')'$, the likelihood function of $\boldsymbol{\theta}$ can be represented as:

$$L(\boldsymbol{\theta}|\boldsymbol{X}_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f\left(\Delta\boldsymbol{X}_i\left(t_{i,j}\right)|\boldsymbol{\beta}_i\right) f_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_i\right) d\beta_i^{(1)} d\beta_i^{(2)} \ldots d\beta_i^{(p)} \tag{38}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}\boldsymbol{\gamma}}, \boldsymbol{\Sigma})$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(p)})'$. $f(\Delta\boldsymbol{X}_i(t_{i,j})|\boldsymbol{\beta}_i)$ represents the conditional PDF of $\Delta\boldsymbol{X}_i(t_{i,j})$ given $\boldsymbol{\beta}_i = (\beta_i^{(1)}, \beta_i^{(2)}, \ldots, \beta_i^{(p)})'$. $f_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_i\right)$ represents the joint PDF of $\boldsymbol{\beta}_i$. On account of that different products are independent, the likelihood function of $\boldsymbol{\theta}$ is thus represented as:

$$L(\boldsymbol{\theta}|\boldsymbol{X}) = \prod_{i=1}^{m} L(\boldsymbol{\theta}|\boldsymbol{X}_i) \tag{39}$$

where $\boldsymbol{X} = (\boldsymbol{X'}_1, \boldsymbol{X'}_2, \ldots, \boldsymbol{X'}_m)'$.

2.7 Degradation Model Updating

Denote the product's degradation measurements at the current time point t_h as $\boldsymbol{X}(t_h) = (X^{(1)}(t_h), X^{(2)}(t_h), \ldots, X^{(p)}(t_h))'$. Now, we investigate the future behavior of $\boldsymbol{X}(t)$ from the current time point, and it is intuitively written by:

$$\boldsymbol{X}(t) = \boldsymbol{X}(t_h) + (\boldsymbol{A}(t) - \boldsymbol{A}(t_h))\boldsymbol{\beta} + \boldsymbol{\xi}_{t,t_h} \tag{40}$$

where $\boldsymbol{\xi}_{t,t_h}$ is normally distributed with mean 0 and variance $(t - t_h)\boldsymbol{\Sigma}$. Considering the latest historical degradation information, a state-space model is utilized to reconstruct Eq. (40), formulated as:

$$\begin{cases} \boldsymbol{\beta}_{(j)} = \boldsymbol{\beta}_{(j-1)} \\ \boldsymbol{X}\left(t_j\right) = \boldsymbol{X}\left(t_{j-1}\right) + \Delta\boldsymbol{A}\left(t_j\right)\boldsymbol{\beta}_{(j-1)} + \boldsymbol{\xi}_{t_j,t_{j-1}} \end{cases} \tag{41}$$

where $t_0 = 0$, $\boldsymbol{X}(t_0) = \boldsymbol{0}$, $\Delta\boldsymbol{A}(t_j) = \boldsymbol{A}(t_j) - \boldsymbol{A}(t_{j-1})$. $\boldsymbol{\xi}_{t_j,t_{j-1}}$ and $\boldsymbol{\beta}_{(0)}$ are normally distributed with mean 0 and $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, variance $t_j - t_{j-1})\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, respectively. Since $\boldsymbol{\beta}_{(j)}$ is depended on the historical degradation information, it is treated as the latent state. Thus, the product's degradation measurements of the target product are rewritten by:

$$\boldsymbol{X}(t) = \boldsymbol{X}(t_h) + (\boldsymbol{A}(t) - \boldsymbol{A}(t_h))\boldsymbol{\beta}_{(h)} + \boldsymbol{\xi}_{t,t_h}$$

Apparently, how to estimate $\boldsymbol{\beta}_{(h)}$ using the historical degradation information becomes a critical issue. To solve this issue, the strong tracking filter (STF) [17] is

applied. Set $\boldsymbol{Y}(t_j) = \boldsymbol{X}(t_j) - \boldsymbol{X}(t_{j-1})$ $(j = 1, 2, \ldots, h)$, the following steps describe the steps of the estimation process.

Step 1: Set $\hat{\boldsymbol{\beta}}_{(0)} = \boldsymbol{\mu}_\beta$, $\boldsymbol{P}_{0|0} = \boldsymbol{\Sigma}_\beta$, α, and κ. Here, α $(\alpha \geq 1)$ and κ $(0 < \kappa \leq 1)$ are the soften factor and the forgetting factor, respectively, and heuristic methods are employed to select proper values [17];

Step 2: Compute the fading factor $\phi(t_j)$ using $\phi(t_j) = \max\{\phi_0, 1\}$, where the subscript j starts from $j = 1$. ϕ_0 is calculated by:

$$\phi_0 = \frac{tr\big(\boldsymbol{N}\big(t_j\big)\big)}{tr\big(\boldsymbol{M}\big(t_j\big)\big)} \tag{42}$$

where $\boldsymbol{N}(t_j) = \boldsymbol{V}_0(t_j) - \alpha(t_j - t_{j-1})\boldsymbol{\Sigma}$, $\boldsymbol{M}(t_j) = \Delta\boldsymbol{A}(t_j)\boldsymbol{P}_{j-1|j-1}\Delta\boldsymbol{A}(t_j)'$. $tr(\mathbf{N})$ denotes the trace of matrix $\mathbf{N}$, and

$$\boldsymbol{V}_0\big(t_j\big) = \begin{cases} \boldsymbol{r}_1\boldsymbol{r}_1' & j = 1 \\ \frac{\kappa \boldsymbol{V}_0\left(t_{j-1}\right) + \boldsymbol{r}_j\boldsymbol{r}_j'}{1+\kappa}, & j > 1 \end{cases}$$

where $\boldsymbol{r}_j = \boldsymbol{Y}(t_j) - \Delta\boldsymbol{A}(t_j)\hat{\boldsymbol{\beta}}_{(j-1)}$;

Step 3: Estimate the latent state $\boldsymbol{\beta}_{(j)}$ by:

$$\hat{\boldsymbol{\beta}}_{(j)} = \hat{\boldsymbol{\beta}}_{(j-1)} + \boldsymbol{K}_j\boldsymbol{r}_j \tag{43}$$

where

$\boldsymbol{K}_j = \boldsymbol{P}_{j|j-1}\Delta\boldsymbol{A}\big(t_j\big)'\Big(\Delta\boldsymbol{A}\big(t_j\big)\boldsymbol{P}_{j-1|j-1}\Delta\boldsymbol{A}\big(t_j\big)' + \big(t_j - t_{j-1}\big)\boldsymbol{\Sigma}\Big)^{-1}\boldsymbol{P}_{j|j-1} = diag\big(\phi\big(t_j\big), \phi\big(t_j\big), \ldots, \phi\big(t_j\big)\big)_{p\times p}\boldsymbol{P}_{j-1|j-1}$;

Step 4: Calculate the covariance of $\boldsymbol{\beta}_{(j)}$ by:

$$\boldsymbol{P}_{j|j} = \big(\boldsymbol{I} - \boldsymbol{K}_j\Delta\boldsymbol{A}\big(t_j\big)\big)\boldsymbol{P}_{j|j-1} \tag{44}$$

where $\boldsymbol{I} = diag(1, 1, \ldots, 1)_{p\times p}$;

Step 5: Set $j = j + 1$, and repeat Steps 2 to Step4 until $j = h$.

As a result, the latent state $\boldsymbol{\beta}_{(h)}$ is estimated using $\boldsymbol{X}_{1:h} = (\boldsymbol{X}(t_1), \boldsymbol{X}(t_2), \ldots, \boldsymbol{X}(t_h))'$, and it approximately obeys a normal distribution with mean $\hat{\boldsymbol{\beta}}_{(h)}$ and covariance matrix $\boldsymbol{P}_{h|h}$.

2.8 Residual Life Prediction

Assume that $X^{(k)}(t)$ is always below the corresponding failure threshold $D^{(k)}$ until time point t_h, the RL of the kth PC, denoted as $L_h^{(k)}$, can be written as:

$$L_h^{(k)} = \inf\left\{l : X^{(k)}(l + t_h) \geq D^{(k)} | \boldsymbol{X}_{1:h}, X^{(k)}\left(t_j\right) < D^{(k)}, j = 1, 2, \ldots, h\right\} \quad (45)$$

Denote the product's RL at time point t_h as L_h, and it can be represented by:

$$L_h = \min\left\{L_h^{(1)}, L_h^{(2)}, \ldots, L_h^{(p)}\right\} \quad (46)$$

Then, the PDF of L_h can be calculated based on $\boldsymbol{\beta}_{(h)}$ and $\boldsymbol{X}_{1:h}$ according to the total law of probability:

$$f_{L_h}(l|\boldsymbol{X}_{1:h}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{L_h}\left(l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) f_{\boldsymbol{\beta}_{(h)}}\left(\boldsymbol{\beta}_{(h)}|\boldsymbol{X}_{1:h}\right) d\beta_{(h)}^{(1)} d\beta_h^{(2)} \ldots d\beta_h^{(p)} \quad (47)$$

where $f_{\boldsymbol{\beta}_{(h)}}(\boldsymbol{\beta}_{(h)}|\boldsymbol{X}_{1:h})$ represents the joint PDF of $\boldsymbol{\beta}_{(h)}$, formulated as:

$$f_{\boldsymbol{\beta}_{(h)}}\left(\boldsymbol{\beta}_{(h)}|\boldsymbol{X}_{1:h}\right) \approx (2\pi)^{-\frac{p}{2}} |\boldsymbol{P}_{h|h}|^{-\frac{1}{2}} \exp\left(-\frac{\left(\boldsymbol{\beta}_{(h)} - \hat{\boldsymbol{\beta}}_{(h)}\right)' \boldsymbol{P}_{h|h}^{-1}\left(\boldsymbol{\beta}_{(h)} - \hat{\boldsymbol{\beta}}_{(h)}\right)}{2}\right)$$

However, the closed-form of $f_{L_h}(l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h})$ is almost impossible to be obtained because the PCs are nonlinear and dependent. In this context, two methods, i.e., the simulation-based method and the approximate method, are employed to tackle this issue.

The core of the simulation-based method is to generate multiple residual lives of the product with multiple PCs by simulation, and the product's RL can then be estimated based on these simulated residual lives. The following steps describe the RL estimation given the product's historical degradation information up to $\boldsymbol{X}_{1:h}$.

Step 1: Estimate $\boldsymbol{\beta}_{(h)}$, and it is expressed as $\boldsymbol{\beta}_{(h)} \sim AN(\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{P}_{h|h})$;

Step 2: Generate M realizations of $\boldsymbol{\beta}_{(h)}$, denoted as $\boldsymbol{\beta}_{(h),i}$ $(i = 1, 2, \ldots, M)$, from $N(\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{P}_{h|h})$;

Step 3: Simulate M possible residual lives based on $\boldsymbol{\beta}_{(h),i}$ $(i = 1, 2, \ldots, M)$. By setting $T_0 = t_h$, $j = 0$, and $\boldsymbol{X}_i(T_0) = \boldsymbol{X}(t_h)$, the simulated degradation measurement value at time point $T_{j+1} = (j + 1)\Delta l + t_h$ can be calculated by:

$$\boldsymbol{X}_i\left(T_{j+1}\right) = \boldsymbol{X}_i\left(T_j\right) + \boldsymbol{\beta}_{(h),i}\left(\boldsymbol{A}\left(T_{j+1}\right) - \boldsymbol{A}\left(T_j\right)\right) + \sqrt{\Delta l}\boldsymbol{W}\boldsymbol{u}_{j+1} \quad (48)$$

where Δl denotes the discretization step, $\boldsymbol{W}$ is a lower triangular matrix that satisfies $\boldsymbol{W}\boldsymbol{W}' = \boldsymbol{\Sigma}$. $\boldsymbol{u}_{j+1}$ is a column random vector generated from $N(\boldsymbol{0}, \boldsymbol{I})$. In Eq. (48), if the condition, i.e., $\exists k \in \{1, 2, \ldots, p\}$ such that $X_i^{(k)}(T_{j+1}) \geq D^{(k)}$, holds, the product's simulated RL can be calculated by $L_{h,i} = (j+1)\Delta l$. Then, another simulated RL can be obtained by setting $i = i + 1 (i < M)$ and repeating Step 3. Otherwise, set $j = j + 1$, the simulated degradation process is rerun until a failure occurs;

Step 4: Calculate the mean value and the PDF of the product's RL at time point t_h. The mean value of the product's RL at time point t_h is estimated using the mean value of $L_{h,i}$ $(i = 1, 2, \ldots, M)$. Moreover, parametric or non-parametric methods are utilized to obtain the PDF of the product's RL using the simulated residual lives.

Generally, it necessary to obtain a closed-form the PDF of the product's RL in real-world applications such as the maintenance arrangement. To this end, parametric distribution models such as the Weibull distribution and the log-normal distribution can be used to fit these simulated residual lives. However, uncertainty exists since the estimation is conducted using simulated residual lives. Thus, the approximate method is employed here.

In accordance to Eq. (46), given $\boldsymbol{\beta}_{(h)}$, one can obtain

$$P\left(L_h \geq l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) = P\left(L_h^{(1)} \geq l, L_h^{(2)} \geq l, \ldots, L_h^{(p)} \geq l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) \tag{49}$$

Set $Y^{(k)}(l) = X^{(k)}(l + t_h) - X^{(k)}(t_h)$, $P(L_h \geq l)$ is rewritten as:

$$\begin{aligned} &P\left(L_h \geq l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) \\ &= P\left(\sup_{s \leq l} Y^{(1)}(s) \leq D_h^{(1)}, \sup_{s \leq l} Y^{(2)}(s) \leq D_h^{(2)}, \ldots, \sup_{s \leq l} Y^{(p)}(s) \leq D_h^{(p)}|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) \end{aligned} \tag{50}$$

Denote the CDF of $\sup_{s \leq l} Y^{(k)}(s)$ as $F_{k,h}(y, l)$. According to Sklar's theorem [18], there exists a unique copula function, denoted as $C(u_1, u_2, \ldots, u_p)$, such that

$$P\left(L_h \geq l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) = C\left(F_{1,h}\left(D_h^{(1)}, l\right), F_{2,h}\left(D_h^{(2)}, l\right), \ldots, F_{p,h}\left(D_h^{(p)}, l\right)\right) \tag{51}$$

The conditional PDF of L_h can then be represented by:

$$\begin{aligned} f_{L_h}\left(l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) \approx \sum_{k=1}^{p} &\frac{\partial C\left(F_{1,h}\left(D_h^{(1)}, l\right), F_{2,h}\left(D_h^{(2)}, l\right), \ldots, F_{k,h}\left(D_h^{(p)}, l\right)\right)}{\partial F_{k,h}} \\ &f_{L_h^{(k)}}\left(l|\beta_{(h)}^{(k)}, \boldsymbol{X}_{1:h}\right) \end{aligned} \tag{52}$$

$f_{L_h}(l|\boldsymbol{X}_{1:h})$ can be further approximately calculated by:

$$f_{L_h}(l|\boldsymbol{X}_{1:h}) \approx \frac{1}{M}\sum_{i=1}^{M} f_{L_h}\left(l|\boldsymbol{\beta}_{(h),i}, \boldsymbol{X}_{1:h}\right) \tag{53}$$

Now, how to obtain the first partial derivative of $C(u_1, u_2, \ldots, u_p)$ becomes a critical issue. Here, a proper approximation method is presented. Suppose that p continuous random variables, denoted as, $X_1, X_2, \ldots, X_p$, and denote their joint CDF as $F(x_1, x_2, \ldots, x_p)$. Let $f_k(x_k)$ and $F_k(x_k)$ respectively represent the PDF and the CDF of X_k. Given $C(u_1, u_2, \ldots, u_p)$ such that $F(x_1, x_2, \ldots, x_p) = C(F_1(x_1), F_2(x_2), \ldots, F_p(x_p))$, one can obtain

$$\begin{aligned} &C\left(u_1, \ldots u_{i-1}, u_{i+1}, \ldots, u_p|u_i\right)|_{u_k=\tilde{u}_k, k=1,2,\ldots,p} \\ &= F\left(x_1, \ldots x_{i-1}, x_{i+1}, \ldots, x_p|x_i\right)|_{x_k=\tilde{x}_k, k=1,2,\ldots,p} \end{aligned} \tag{54}$$

where $\tilde{x}_k = F_k^{-1}(\tilde{u}_k)$ $(k = 1, 2, \ldots, p)$.

Set $\boldsymbol{Y}_{\sup,-k} = (\sup_{s\le l} Y^{(1)}(s), \ldots, \sup_{s\le l} Y^{(k-1)}(s), \sup_{s\le l} Y^{(k+1)}(s), \ldots, \sup_{s\le l} Y^{(p)}(s))'$. Thus, the first partial derivative of $C(u_1, u_2, \ldots, u_p)$ with respect to u_k can be estimated using the conditional CDF of $\boldsymbol{Y}_{\sup,-k}$ given $\sup_{s\le l} Y^{(k)}(s)$ and $\boldsymbol{\beta}_{(h)}$. Set $\boldsymbol{Y}_{-k} = (Y^{(1)}(l), \ldots, Y^{(k-1)}(l), Y^{(k+1)}(l), \ldots, Y^{(p)}(l))'$, thus, $F_{Y_{-k}|Y^{(k)}(l)}(y^{(1)}, \cdots, y^{(k-1)}, y^{(k+1)}, \ldots, y^{(p)}|Y^{(k)}(l) = y^{(k)})$, the conditional CDF of $\boldsymbol{Y}_{-k}$ given $Y^{(k)}(l)$ and $\boldsymbol{\beta}_{(h)}$ is used to approximate the value of the conditional CDF of $\boldsymbol{Y}_{\sup,-k}$. One can then obtain

$$\begin{aligned} &f_{L_h}\left(l|\boldsymbol{\beta}_{(h)}, \boldsymbol{X}_{1:h}\right) \\ &\approx \sum_{k=1}^{p} F_{\boldsymbol{Y}_{-k}|Y^{(k)}(l)}\left(D_h^{(1)}, \ldots, D_h^{(k-1)}, D_h^{(k+1)}, \ldots, D_h^{(p)}|Y^{(k)}(l) = D_h^{(k)}\right) \\ &\quad f_{L_h^{(k)}}\left(l|\beta_{(h)}^{(k)}, \boldsymbol{X}_{1:h}\right) \end{aligned} \tag{55}$$

Set $\boldsymbol{Y} = (Y^{(1)}(l), Y^{(2)}(l), \ldots, Y^{(p)}(l))'$. Given $\boldsymbol{\beta}_{(h)}$, one can obtain

$$\boldsymbol{Y}|\boldsymbol{\beta}_{(h)} \sim N\left((\boldsymbol{A}(l + t_h) - \boldsymbol{A}(t_h))\boldsymbol{\beta}_{(h)}, l\boldsymbol{\Sigma}\right)$$

Thus, given $Y^{(k)}(l)$ and $\boldsymbol{\beta}_{(h)}$, $\boldsymbol{Y}_{-k}$ also obeys the normal distribution.

Since $\sup_{s\le l} Y^{(k)}(s) = Y^{(k)}(l)$ $(k = 1, 2, \ldots, p)$, it is reasonable to approximate the conditional CDF of $\boldsymbol{Y}_{\sup,-k}$ using the conditional CDF of $\boldsymbol{Y}_{-k}$ for non-decreasing degradation processes. Here, a Wiener process with a nonlinear drift is used to characterize the evolution of $Y^{(k)}(l)$. In this case, the degradation rate over time is accelerated. When l takes a large value or $\sigma^{(k)}$ $(k = 1, 2, \ldots, p)$ takes a small value, the probability of the degradation increment in a time interval that is less than zero

will be close to zero. Therefore, $Y^{(k)}(l)$ gets close to $\sup_{s \le l} Y^{(k)}(s)$ $(k = 1, 2, \ldots, p)$, leading to a reasonable approximation.

2.9 Parameters Estimation

Here, we use the expectation–maximization (EM) algorithm solve the parameter estimation problem where $\boldsymbol{\beta}_i$ $(i = 1, 2, \ldots, m)$ is treated as the latent variable. Specifically, two steps, i.e., the E-step and M-step, are iteratively conducted until the terminal condition is reached.

E-step: Calculate the expectation of the log-likelihood function of $\boldsymbol{\beta}_i$ $(i = 1, 2, \ldots, m)$ based on the current value of $\boldsymbol{\theta}$ and the observed degradation data $\boldsymbol{X}$;

M-step: Update $\boldsymbol{\theta}$ by maximizing the expectation of the log-likelihood function result of the E-step.

Given the random effects $\boldsymbol{\beta}_i$ $(i = 1, 2, \ldots, m)$ and the complete data $\boldsymbol{Z} = (\boldsymbol{X}', \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m)'$. Then, the complete log-likelihood function of $\boldsymbol{\theta}$ can be written as:

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{Z}) = \sum_{i=1}^{m} \Bigg\{ & -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\beta| - \frac{1}{2}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}_\beta\right)'\left(\boldsymbol{\Sigma}_\beta\right)^{-1}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}_\beta\right) \\ & - \frac{n_i p}{2}\ln(2\pi) - \frac{1}{2}\sum_{j=1}^{n_i}\ln|\Delta t_{i,j}\boldsymbol{\Sigma}| \\ & - \frac{1}{2}\sum_{j=1}^{n_i}\left(\Delta \boldsymbol{X}_i(t_{i,j}) - \Delta \boldsymbol{A}(t_{i,j})\boldsymbol{\beta}_i\right)'\left(\Delta t_{i,j}\boldsymbol{\Sigma}\right)^{-1}\left(\Delta \boldsymbol{X}_i(t_{i,j}) - \Delta \boldsymbol{A}(t_{i,j})\boldsymbol{\beta}_i\right) \Bigg\}. \end{aligned} \tag{56}$$

Denote $\boldsymbol{\theta}^{(I)}$, $\boldsymbol{\mu}_\beta^{(I)}$, $\boldsymbol{\Sigma}_\beta^{(I)}$, $\boldsymbol{\gamma}^{(I)}$, and $\boldsymbol{\Sigma}^{(I)}$ as the current values of $\boldsymbol{\theta}$, $\boldsymbol{\mu}_\beta$, $\boldsymbol{\Sigma}_\beta$, $\boldsymbol{\gamma}$, and $\boldsymbol{\Sigma}$ after I iterations, respectively. Thus,

$$\left(\boldsymbol{\beta}_i|\boldsymbol{\theta}^{(I)}, \boldsymbol{X}\right) \sim N\left(\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)}\right) \tag{57}$$

where

$$\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} = \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)}\left(\sum_{j=1}^{n_i}\Delta \boldsymbol{X}_i(t_{i,j})'\left(\Delta t_{i,j}\boldsymbol{\Sigma}^{(I)}\right)^{-1}\Delta \boldsymbol{A}_{i,j}^{(I)}(t_{i,j}) + \boldsymbol{\mu}_\beta^{(I)\prime}\left(\boldsymbol{\Sigma}_\beta^{(I)}\right)^{-1}\right)' \tag{58}$$

and

$$\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)} = \left(\sum_{j=1}^{n_i} \Delta \boldsymbol{A}^{(I)}(t_{i,j}) \left(\Delta t_{i,j} \boldsymbol{\Sigma}^{(I)} \right)^{-1} \Delta \boldsymbol{A}^{(I)}(t_{i,j}) + \left(\boldsymbol{\Sigma}_{\beta}^{(I)} \right)^{-1} \right)^{-1} \tag{59}$$

The expectation of $l(\boldsymbol{\theta}|\boldsymbol{Z})$ at the $(I+1)$th iteration can thus be represented by:

$$\begin{aligned} E\left(l\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(I)}, \boldsymbol{X}\right)\right) = \sum_{i=1}^{m} \Bigg\{ & -\frac{(n_i+1)p}{2} \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_{\beta}| \\ & - \frac{1}{2} \left(tr\left(\left(\boldsymbol{\Sigma}_{\beta}\right)^{-1} \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)} \right) + \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} \left(\boldsymbol{\Sigma}_{\beta}\right)^{-1} \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} - 2\boldsymbol{\mu}_{\beta}' \left(\boldsymbol{\Sigma}_{\beta}\right)^{-1} \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} + \boldsymbol{\mu}_{\beta}' \left(\boldsymbol{\Sigma}_{\beta}\right)^{-1} \boldsymbol{\mu}_{\beta} \right) \\ & - \frac{1}{2} \sum_{j=1}^{n_i} \Big\{ \ln |\Delta t_{i,j} \boldsymbol{\Sigma}| + \Delta \boldsymbol{X}_i(t_{i,j})' \left(\Delta t_{i,j} \boldsymbol{\Sigma}\right)^{-1} \Delta \boldsymbol{X}_i(t_{i,j}) \\ & - 2\Delta \boldsymbol{X}_i(t_{i,j})' \left(\Delta t_{i,j} \boldsymbol{\Sigma}\right)^{-1} \Delta \boldsymbol{A}(t_{i,j}) \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} \\ & + tr\left(\Delta \boldsymbol{A}(t_{i,j}) \left(\Delta t_{i,j} \boldsymbol{\Sigma}\right)^{-1} \Delta \boldsymbol{A}(t_{i,j}) \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)} \right) \\ & + \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} \Delta \boldsymbol{A}(t_{i,j}) \left(\Delta t_{i,j} \boldsymbol{\Sigma}\right)^{-1} \Delta \boldsymbol{A}(t_{i,j}) \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} \Big\} \Bigg\}, \end{aligned} \tag{60}$$

In the M-step, the following equations are obtained:

$$\frac{\partial E(l(\boldsymbol{\theta}|\boldsymbol{X}))}{\partial \boldsymbol{\mu}_{\beta}} = \sum_{i=1}^{m} \left\{ \left(\boldsymbol{\Sigma}_{\beta}\right)^{-1} \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} - \left(\boldsymbol{\Sigma}_{\beta}\right)^{-1} \boldsymbol{\mu}_{\beta} \right\} = \boldsymbol{0} \tag{61}$$

$$\frac{\partial E\left(l\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(I)}, \boldsymbol{X}\right)\right)}{\partial \left(\boldsymbol{\Sigma}_{\beta}\right)^{-1}} = \frac{1}{2} \sum_{i=1}^{m} \left\{ \boldsymbol{\Sigma}_{\beta} - \left(\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)} + \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} - 2\boldsymbol{\mu}_{\beta} \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} + \boldsymbol{\mu}_{\beta} \boldsymbol{\mu}_{\beta}' \right) \right\} = \boldsymbol{0} \tag{62}$$

$$\begin{aligned} \frac{\partial E(l(\boldsymbol{\theta}|\boldsymbol{X}))}{\partial (\boldsymbol{\Sigma})^{-1}} = \frac{1}{2} \sum_{i=1}^{m} \Bigg\{ & n_i \boldsymbol{\Sigma} - \sum_{j=1}^{n_i} \frac{1}{\Delta t_{i,j}} \Big\{ \Delta \boldsymbol{X}_i(t_{i,j}) \Delta \boldsymbol{X}_i(t_{i,j})' \\ & - 2\Delta \boldsymbol{A}(t_{i,j}) \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} \Delta \boldsymbol{X}_i(t_{i,j})' \\ & + \Delta \boldsymbol{A}(t_{i,j}) \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)} \Delta \boldsymbol{A}(t_{i,j}) + \Delta \boldsymbol{A}(t_{i,j}) \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)} \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} \Delta \boldsymbol{A}(t_{i,j}) \Big\} \Bigg\} = \boldsymbol{0}. \end{aligned} \tag{63}$$

Therefore, the iterative formulas of $\boldsymbol{\mu}_{\beta}^{(I+1)}$, $\boldsymbol{\Sigma}_{\beta}^{(I+1)}$, and $\boldsymbol{\Sigma}^{(I+1)}$ can be expressed as:

$$\boldsymbol{\mu}_{\beta}^{(I+1)} = \frac{\sum_{i=1}^{m} \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)}}{m} \tag{64}$$

$$\boldsymbol{\Sigma}_{\beta}^{(I+1)} = \frac{1}{m}\sum_{i=1}^{m}\left(\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)} + \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)}\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} - 2\boldsymbol{\mu}_{\beta}^{(I+1)}\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime} + \boldsymbol{\mu}_{\beta}^{(I+1)}\boldsymbol{\mu}_{\beta}^{(I+1)\prime}\right) \tag{65}$$

$$\begin{aligned}\boldsymbol{\Sigma}^{(I+1)} = \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\Delta t_{i,j}}\Big\{&\Delta \boldsymbol{X}_i\left(t_{i,j}\right)\Delta \boldsymbol{X}_i\left(t_{i,j}\right)' - 2\Delta \boldsymbol{A}\left(t_{i,j}\right)\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)}\Delta \boldsymbol{X}_i\left(t_{i,j}\right)\prime \\ &+\Delta \boldsymbol{A}\left(t_{i,j}\right)\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)}\Delta \boldsymbol{A}\left(t_{i,j}\right) + \Delta \boldsymbol{A}\left(t_{i,j}\right)\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)}\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)\prime}\Delta \boldsymbol{A}\left(t_{i,j}\right)\Big\}.\end{aligned} \tag{66}$$

Now, $E(l(\boldsymbol{\theta}|\boldsymbol{\theta}^{(I)}, \boldsymbol{X}))$ can be substituted with $\boldsymbol{\mu}_{\beta}^{(I+1)}$, $\boldsymbol{\Sigma}_{\beta}^{(I+1)}$, and $\boldsymbol{\Sigma}^{(I+1)}$. Denote the corresponding function of $\boldsymbol{\gamma}$ as $Q(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(I)}, \boldsymbol{\mu}_{\beta}^{(I+1)}, \boldsymbol{\Sigma}_{\beta}^{(I+1)}, \boldsymbol{\Sigma}^{(I+1)}, \boldsymbol{X})$. By maximizing $Q(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(I)}, \boldsymbol{\mu}_{\beta}^{(I+1)}, \boldsymbol{\Sigma}_{\beta}^{(I+1)}, \boldsymbol{\Sigma}^{(I+1)}, \boldsymbol{X})$, a new value of $\boldsymbol{\gamma}$, denoted by $\boldsymbol{\gamma}^{(I+1)}$, can be obtained. Here, the MATLAB function "fminsearch" is used to conduct the multi-dimensional search of $\boldsymbol{\gamma}$.. Finally, by substituting $\boldsymbol{\gamma}^{(I+1)}$ into Eq. (66), the value of $\boldsymbol{\Sigma}^{(I+1)}$ can be obtained. The entire procedures can be summarized as follows.

Step 1: Initialize $\boldsymbol{\mu}_{\beta}^{(0)}$, $\boldsymbol{\Sigma}_{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, and $\boldsymbol{\Sigma}^{(0)}$. Details are provided in the Appendix;

Step 2: Compute $\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_i}^{(I)}$ and $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{(I)}$ $(i = 1, 2, \ldots, m)$ based on $\boldsymbol{\mu}_{\beta}^{(I)}$, $\boldsymbol{\Sigma}_{\beta}^{(I)}$, $\boldsymbol{\gamma}^{(I)}$, and $\boldsymbol{\Sigma}^{(I)}$;

Step 3: Calculate $\boldsymbol{\mu}_{\beta}^{(I+1)}$, $\boldsymbol{\Sigma}_{\beta}^{(I+1)}$, $\boldsymbol{\gamma}^{(I+1)}$, and $\boldsymbol{\Sigma}^{(I+1)}$;

Step 4: Repeat Step 2 to Step 3 until the deviation of the expectations of the log-likelihood function in the two adjacent iterations is less than a predefined threshold.

3 Illustrative Examples

In this section, two numerical examples are set to verify the effectiveness of the proposed methods.

3.1 Illustrative Example for the Univariate Nonlinear Wiener Process

The original degradation data is collected from [19] where 21 products with initial crack length of 0.9 inches are tested for fatigue crack development. When the cumulative crack length exceeds 0.7 inches (or the crack length crosses over 1.6 inches), the product is regarded as failed. The measured frequency is 0.01 millions of cycles. Here, the degradation data for each product with terminated time of 0.09 millions of cycles are used. ξ is set as 0.01, and the simulated measurement errors are added to the original degradation data with respect to each measurement time. The simulated degradation measurements are shown in Fig. 1.

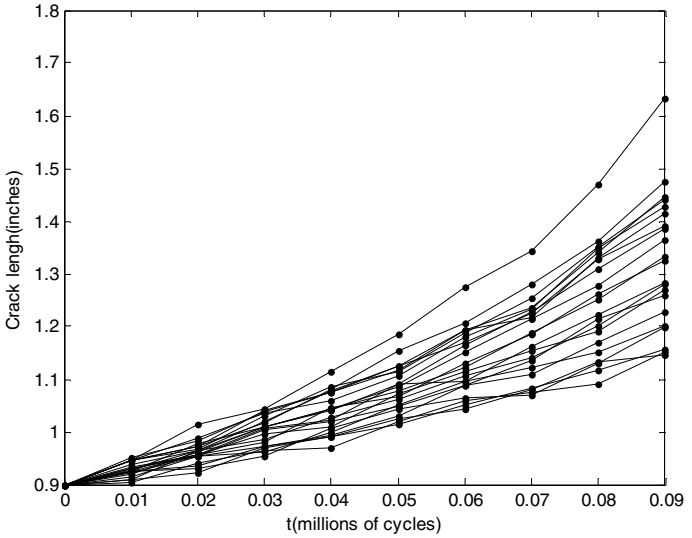


Fig. 1 The measurements of crack length over time

To make use of the proposed degradation model, a transformation of substituting $Y(t)$ with $Y(0)$ is conducted on the degradation data. It is observed that nonlinearity exists in the degradation paths. Therefore, a nonlinear degradation model is employed such that the dynamics of the degradation can be captured. As investigated in [20], the mean deterioration is often in proportion to time with a power law, namely, $E(X(t)) \propto t^b$. With respect to model M_2, $E(X(t)|a) = a\Lambda(t;\boldsymbol{\theta})$. As a result, $\Lambda(t;\boldsymbol{\theta})$ and $\tau(t;\boldsymbol{\gamma})$ are assumed to be t^b and t^γ, respectively. Notably, it is readily to extend to other forms of $\Lambda(t;\boldsymbol{\theta})$ and $\tau(t;\boldsymbol{\gamma})$.

We treat the transformed degradation data as the population-based degradation data and use both M_2 and M_1 to fit it. The log-LF under model M_2 is 483.9778 and that under model M_1 is 480.0056, respectively. Their AIC values are $-$ 955.9556 and $-$ 950.0112, respectively, indicating that model M_2 fits the data better than model M_1. Table 1 lists the estimates of parameters of model M_2, including the point estimate, standard deviation (Std), and 95% confidence interval.

Table 1 Parameter estimation results under model M_2

Para	Estimate	Std	2.5%	97.5%
μ_0	10.1219	0.9246	8.3098	11.9341
σ_0	2.9792	0.4791	2.0401	3.9182
σ	0.3229	0.0519	0.2211	0.4247
ξ	0.0088	0.0014	0.0060	0.0116
b	1.3216	0.0238	1.2749	1.3683
γ	1.7722	0.1689	1.4411	2.1033

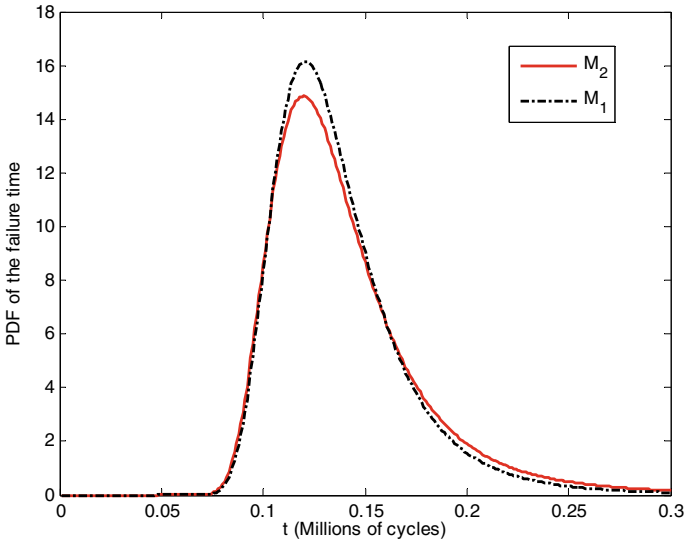


Fig. 2 PDFs of the failure time based on model M_1 and M_2 (fatigue cracks)

Based on the parameter estimation results, the PDFs of the failure time based on model M_1 and M_2 can be calculated, and they are plotted in Fig. 2, where some distinct differences exist between the two PDF curves.

Now, the product's degradation path can be generated, as shown in Fig. 3, via simulation with sampling interval of 0.0001 millions of cycles. For comparison, the true value of degradation, denoted as $X(t)$, is also presented. Intuitively, the product's failure time is 0.1087 millions of cycles.

Figure 4 shows the product's RL estimation results at different time points (i.e., $\boldsymbol{t}_{1:10} = (0.01, 0.02, \ldots, 0.1)$). In Fig. 4, the actual RL and the estimated RL at each time point are marked with different shapes. The corresponding 95% confidence interval of the estimated RL is listed in Table 2. Apparently, the actual RL always falls within the range of the estimated PDF at each time point. Moreover, the variance of RL exhibits a decreasing trend as the time increases.

In the proposed method, the population-based degradation data is utilized to estimate the value of $\boldsymbol{\Theta}_1$. Here, a sensitivity analysis is conducted regarding the relative errors with respect to $\boldsymbol{\Theta}_1$ is investigated. Specifically, the value of θ_l is multiplied by a positive factor, denoted as ρ. Meanwhile, other parameters are unchanged. For illustration purpose, we assume that the actual failure time of product s is T_s. In addition, denote $t_{s,j}$ as the cumulative operating time at measurement epoch j and $l_{s,j}$ the predicted RL at time $t_{s,j}$ ($j = 1, 2, \ldots, I_s$). Thus, the relative error is defined as:

$$RE_s = \frac{1}{I_s} \sum_{j=1}^{I_s} \frac{|t_{s,j} + l_{s,j} - T_s|}{T_s} \tag{67}$$

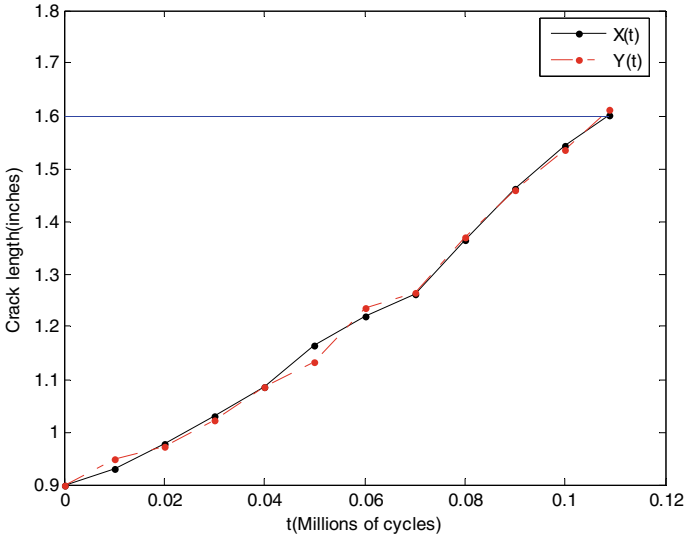


Fig. 3 The product' degradation path

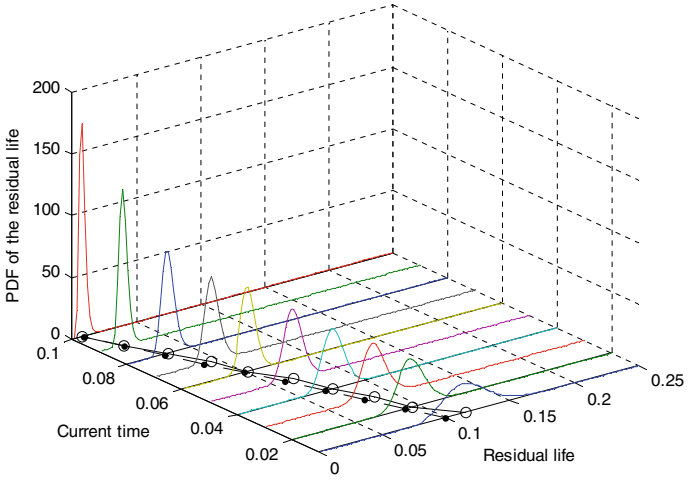


Fig. 4 The product's RL estimation results

Subsequently, the relative errors under various settings of ρ, i.e., $\{0.9, 0.95, 1.05, 1.1\}$, are calculated. Table 3 lists the corresponding sensitivity analysis results. It is observed that when there is a moderate departure from the estimate of $\boldsymbol{\Theta}_1$, the relative error is robust.

Table 2 Point estimate and 95% confidence interval of the estimated RL

t_h	True	Mean	2.5%	97.5%
0.01	0.0987	0.1154	0.0889	0.1559
0.02	0.0887	0.0951	0.0777	0.1193
0.03	0.0787	0.0865	0.0699	0.1054
0.04	0.0687	0.0752	0.0625	0.0947
0.05	0.0587	0.0664	0.0534	0.0805
0.06	0.0487	0.0519	0.0420	0.0658
0.07	0.0387	0.0448	0.0348	0.0586
0.08	0.0287	0.0318	0.0237	0.0436
0.09	0.0187	0.0187	0.0134	0.0252
0.10	0.0087	0.0080	0.0043	0.0131

Table 3 Sensitivity analysis results of relative errors with respect to $\boldsymbol{\Theta}_1$

Para	ρ			
	0.9	0.95	1.05	1.1
μ_0	0.0839	0.0722	0.0508	0.0409
σ_0	0.0631	0.0619	0.0601	0.0594
σ	0.0592	0.0605	0.0618	0.0626
ξ	0.0598	0.0603	0.0615	0.0621
b	0.0670	0.0576	0.0742	0.0979
γ	0.0682	0.0641	0.0582	0.0561

3.2 Illustrative Example for the Multivariate Nonlinear Wiener Process

Motivated by the works in [10, 11], a product is assumed to have three possible fatigue crack positions with initial crack length of 0.9 inches, and the corresponding crack lengths increasing over time are respectively represented as PC1, PC2, and PC3. In addition, the failure threshold with respect to PC1, PC2, and PC3 are respectively set as 1.25, 1.45, and 1.61 inches. For illustration purpose, the dataset in [19] is taken and used under these parameter settings. To conduct the experiments, the original data with terminated time of 0.09 millions of cycles is artificially categorized into three groups, representing the degradation measurements from a PC1, PC2, and PC3, respectively. The grouped data are viewed as the degradation measurements of 7 products with three PCs, as shown in Fig. 5. Clearly, nonlinearity exists in the degradation of each PC. Therefore, a nonlinear structure is appropriate and should be considered. In the proposed degradation model, the nonlinearity involved in the degradation process of the kth PC is characterized by $\Lambda^{(k)}(t; \boldsymbol{\gamma}^{(k)})$. Without loss of generality, $\Lambda^{(k)}(t; \boldsymbol{\gamma}^{(k)})$ is set to be t^{γ_k} $(k = 1, 2, \ldots, p)$.

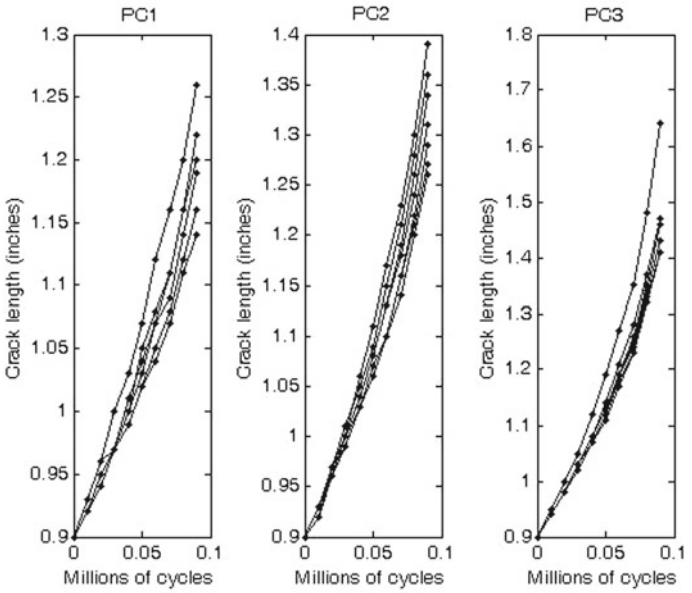


Fig. 5 The development of crack sizes over time

We treat the data in Fig. 5 as the population-based degradation data and estimate the unknown parameters, and they are:

$$\boldsymbol{\mu}_{\beta} = (6.5558, 12.4143, 18.8090)', \boldsymbol{\gamma} = (1.2977, 1.4122, 1.4606)',$$

$$\boldsymbol{\Sigma}_{\beta} = \begin{pmatrix} 0.6172 & 0.9896 & 1.7226 \\ 0.9896 & 1.5895 & 2.7254 \\ 1.7226 & 2.7254 & 5.4640 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 0.0055 & 0.0018 & 0.0045 \\ 0.0018 & 0.0082 & 0.0069 \\ 0.0045 & 0.0069 & 0.0213 \end{pmatrix}$$

Notably, the estimating results of γ_k $(k = 1, 2, 3)$ confirm the embedded nonlinearity. For illustration purpose, these estimated results will be regarded as the ground truth for result comparison. Given $\boldsymbol{\beta}$, the correlation coefficient matrix of $\boldsymbol{X}(t)$ is expressed as:

$$Cor = \begin{pmatrix} 1 & 0.2680 & 0.4158 \\ 0.2680 & 1 & 0.5221 \\ 0.4158 & 0.5221 & 1 \end{pmatrix}.$$

Simulation process is conducted to generate the product's degradation measurements. In order to obtain an exact product's failure time, the simulation process proceeds until a failure occurs. Figure 6 illustrates the product's simulated degradation path. It is observed in Fig. 6 that a failure occurs at the 0.0999 millions of cycles (the discretization step is set to be 0.0001 millions of cycles) and the main cause is related to PC_1. For demonstration purpose, we choose measurement time

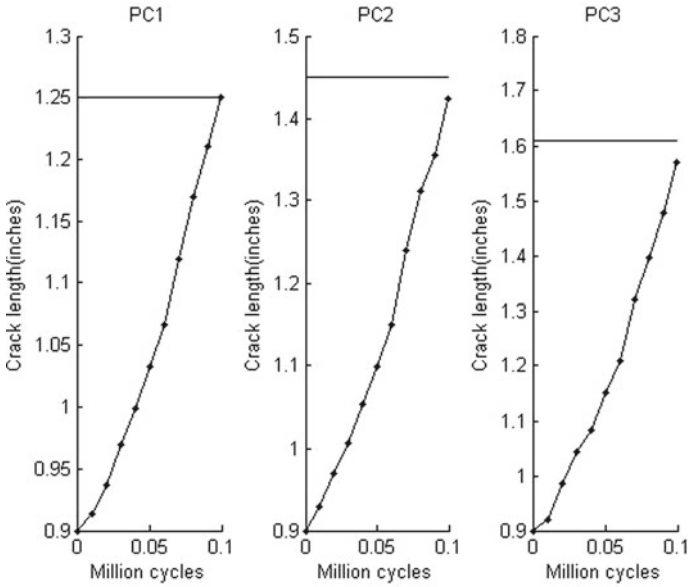


Fig. 6 Simulated degradation path of the target product over time

points $\boldsymbol{t}_{1:9} = \{0.01, 0.02, \ldots, 0.09\}$ and set the current time t_h ($h = 1, 2, \ldots, 9$) to estimate the product's RL.

Here, both a simulation-based method and an approximated method are presented to estimate the product's RL. Specifically, the Weibull distribution with the scale parameter as 0.0835 and the shape parameter as 10.2883 is used in the simulation-based method to fit those simulated residual lives. The RL estimation results at time $t_h = 0.02$ under both two models are calculated and the results are depicted in Fig. 7. Apparently, the approximated method has a better performance. Therefore, it will be used to estimate the product's RL in the following part.

Figure 8 presents the PDFs, means, CDFs of the product's RL. Intuitively, both the PDFs and the CDFs have a regular change at each measurement time point. Particularly, the variance of the PDF of the product's RL decreases as the degradation data accumulates. Moreover, the actual RL always falls within the range of the estimated PDF at each measurement time point. The mean of the estimated RL is close to the real RL and the maximal deviation is merely 0.0042. This value, however, will arise to 0.0095 if the dependency between PCs is ignored.

Now, we can conclude that the proposed method achieves a good performance in estimating the product's RL in such parameter settings. A sensitivity analysis is conducted regarding the relative error with respect to each parameter of the proposed degradation model is investigated. Specifically, the value of θ is multiplied by a positive factor, denoted as ρ, and the value of which is selected from

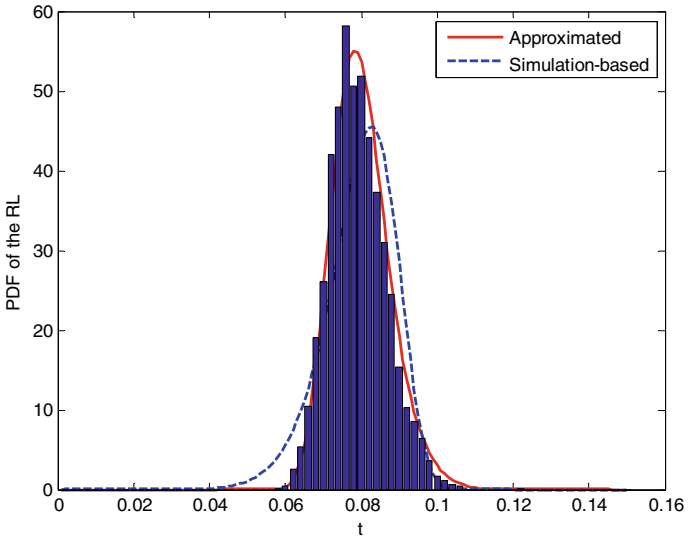


Fig. 7 PDFs of the estimated RL at time 0.02 under two methods

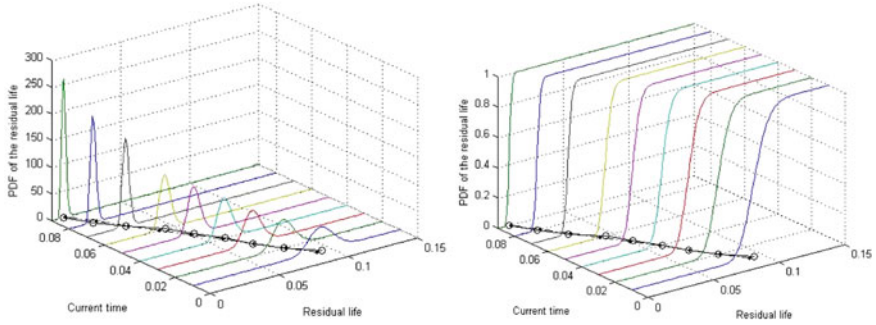


Fig. 8 The PDFs, means, and CDFs of the product's RL at each measurement time point

$\{0.9, 0.95, 1.05, 1.1\}$. Meanwhile, other parameters are unchanged. Here, H and T are set to be 9 and 0.0999, respectively. Table 4 lists the sensitivity analysis results, where ν_k $(k = 1, 2, 3)$ is the standard deviation of $\beta^{(k)}$. It is observed that when there is a moderate departure from the estimate of each parameter, the relative error remains stable.

In order to conduct a further illustration, the degradation paths of 50 products are simulated based on the proposed degradation model. With respect to each product, its RLs are firstly estimated at time points $\boldsymbol{t}_{1:8} = \{0.01, 0.02, \ldots, 0.08\}$ and the relative error is then calculated. If a product failed before 0.08 millions of cycles,

Table 4 The sensitivity analysis results of the relative error with respect to the model parameters

Para	ρ			
	0.9	0.95	1.05	1.1
μ_1	0.02152	0.02028	0.01929	0.02048
μ_2	0.01961	0.02152	0.01949	0.01724
μ_3	0.02473	0.02403	0.01832	0.01879
ν_1	0.01972	0.02132	0.02101	0.02013
ν_2	0.02142	0.02078	0.02014	0.02029
ν_3	0.01894	0.02016	0.02139	0.02201
$\sigma^{(1)}$	0.02064	0.01992	0.01862	0.02086
$\sigma^{(2)}$	0.01967	0.02135	0.02166	0.01987
$\sigma^{(3)}$	0.02154	0.02029	0.01895	0.02044
$\rho_{1,2}$	0.01964	0.01944	0.01964	0.01947
$\rho_{1,3}$	0.02027	0.02001	0.01991	0.02084
$\rho_{2,3}$	0.02087	0.02001	0.02091	0.01987
γ_1	0.02628	0.01934	0.02192	0.01923
γ_2	0.03481	0.02065	0.02099	0.02548
γ_3	0.04113	0.02619	0.03028	0.03651

it is discarded and the simulation process runs again until the terminal condition is satisfied. Here, the relative error of a product based on the RL estimates at $\boldsymbol{t}_{1:4}$ is also evaluated, and the results are shown in Fig. 9. Except in the case of few products, more degradation information can induce a smaller relative error. Inspired by this phenomenon, it is better to update estimation results in a real-time manner so as to obtain an accurate estimation result.

Figure 10 shows the comparison results of relative errors in the dependent and independent cases. The results indicate that, for most of the products, ignoring the dependency between PCs can lead to a larger relative error, emphasizing the significance of taking account of the dependency between PCs.

4 Summary

In this work, we focus on nonlinear degradation modeling and residual life prediction by using Wiener process, both univariate nonlinear and multivariate nonlinear situations are addressed in detail. In the univariate nonlinear scenario, a generalized Wiener process-based degradation model with measurement errors is developed. The proposed model can subsume many existing Wiener process-based models.

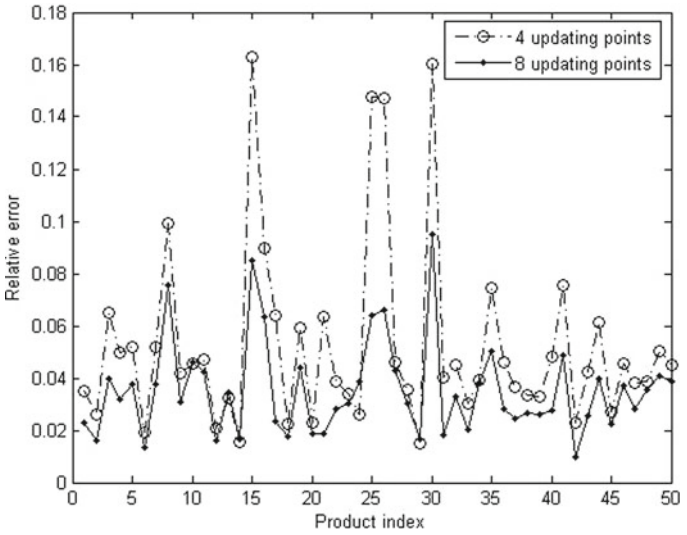


Fig. 9 Relative errors under RL estimates at $\boldsymbol{t}_{1:4}$ and $\boldsymbol{t}_{1:8}$

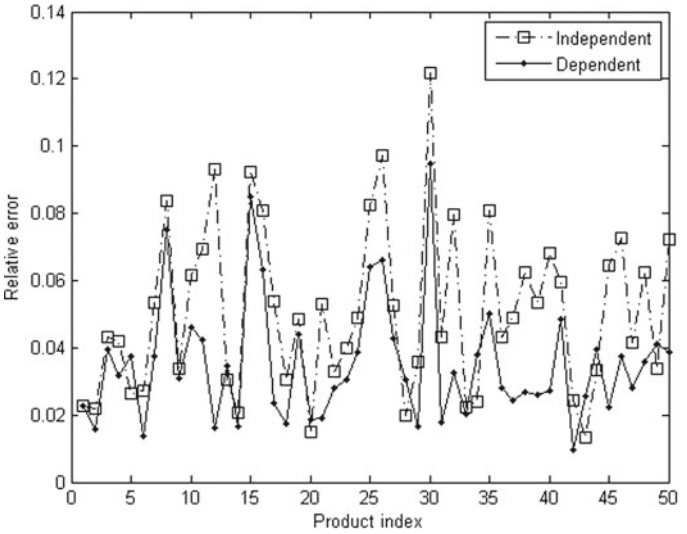


Fig. 10 Relative errors in the dependent and independent cases

Moreover, for meeting the practical requirements in maintenance decision, a closed-form of the PDF of the product's RL can be approximately obtained. With respect to the multivariate nonlinear scenario, a multivariate Wiener degradation process model with nonlinear drifts is given and explored to estimate the product's RL. The numerical examples verities the effectiveness of the proposed method.

References

1. Escobar LA, Meeker WQ (2006) A review of accelerated test models. Stat Sci 21:552–577
2. Si XS, Wang W, Hu CH, Zhou DH (2011) Remaining useful life estimation—a review on the statistical data driven approaches. Eur J Oper Res 213(1):1–14
3. Gebraeel NZ, Lawley MA, Li R, Ryan JK (2005) Residual-life distributions from component degradation signals: a Bayesian approach. IIE Trans 37(6):543–557
4. Long B, Xian W, Jiang L, Liu Z (2013) An improved autoregressive model by particle swarm optimization for prognostics of lithium-ion batteries. Microelectron Reliab 53(6):821–831
5. Wang W, Carr M, Xu W, Kobbacy K (2011) A model for residual life prediction based on Brownian motion with an adaptive drift. Microelectron Reliab 51(2):285–293
6. Liu J, Li Q, Chen W, Yan Y, Qiu Y, Cao T (2019) Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. Int J Hydrogen Energy 44(11):5470–5480
7. Wang X, Balakrishnan N, Guo B (2014) Residual life estimation based on a generalized Wiener degradation process. Reliab Eng Syst Saf 124:13–23
8. Si XS, Wang W, Hu CH, Chen MY, Zhou DH (2013) A Wiener-process-based degradation model with a recursive filter algorithm for remaining useful life estimation. Mech Syst Signal Process 35(1–2):219–237
9. Li N, Lei Y, Yan T, Li N, Han T (2018) A Wiener-process-model-based method for remaining useful life prediction considering unit-to-unit variability. IEEE Trans Industr Electron 66(3):2092–2101
10. Pan Z, Balakrishnan N, Sun Q, Zhou J (2013) Bivariate degradation analysis of products based on Wiener processes and copulas. J Stat Comput Simul 83(7):1316–1329
11. Sari JK, Newby MJ, Brombacher AC, Tang LC (2009) Bivariate constant stress degradation model: LED lighting system reliability estimation with two-stage modelling. Qual Reliab Eng Int 25(8):1067–1084
12. Pan Z, Balakrishnan N, Sun Q (2011) Bivariate constant-stress accelerated degradation model and inference. Commun Stat—Simul Comput 40(2), 247–257
13. Wang X, Balakrishnan N, Guo B (2015) Residual life estimation based on nonlinear-multivariate Wiener processes. J Stat Comput Simul 85(9):1742–1764
14. Whitmore GA (1995) Estimating degradation by a Wiener diffusion process subject to measurement error. Lifetime Data Anal 1(3):307–319
15. Peng CY, Tseng ST (2009) Mis-specification analysis of linear degradation models. IEEE Trans Reliab 58(3):444–455
16. Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1:54–75
17. Zhou DH, Frank PM (1996) Strong tracking filtering of nonlinear time-varying stochastic systems with coloured noise: Application to parameter estimation and empirical robustness analysis. Int J Control 65(2):295–307
18. Nelsen RB (2007) An introduction to copulas. Springer Science & Business Media, New York
19. William QM, Escobar LA (1998) Statistical methods for reliability data. Wiley, New York
20. van Noortwijk JM (2009) A survey of the application of gamma processes in maintenance. Reliab Eng Syst Saf 94(1):2–21

System Reliability Models with Dependent Degradation Processes

Zhanhang Li, Chenyu Han, and David W. Coit

Abstract Interest and associated research for reliability and health prediction and maintenance of infrastructure and industrial products have increased continuously. The study of reliability and health prognosis has become an indispensable field in the overall design and evaluation of systems, industrial products and engineering projects. Previously, the common approaches and mathematical models to describe the condition of products were usually based on the statistical lifetime distribution of the target production. The lifetime distribution is obtained based on the observation and analysis of large quantities of components. However, when it comes to a single component, it can only quantify whether the component is functioning or not, rather than the detailed working condition or deterioration behavior. Therefore, degradation models are introduced to quantify the health conditions of the component based on time dependent observations. Alternatively, on the basis of the degradation model, by introducing the degradation threshold of product failure, the reliability model and the remaining useful life of the product and the corresponding maintenance strategy can also be derived. In practice, the evaluation of the degradation behavior of the system often needs to introduce multiple degradation processes while modeling, and these degradation processes are not always independent of each other. Due to factors inherent in the system or from the external environment, these degradation processes often affect each other and show some commonalities. Examples of such degradation include LED lighting systems (Sari et al. in Qual Reliab Eng Int 25:1067–1084, 2009), operating data of heavy-duty machine tools (Mi et al. in Reliab Eng Syst Saf 174:71–81, 2018), fatigue cracks of two terminals of an electronic device (Rodríguez-Picón et al. in Appl Stoch Model Bus Ind 35:504–521, 2019), etc. In this chapter, we will introduce various degradation models, as well as modeling approaches and reliability analysis to study dependent processes, such as dependent Markov chains, shared shock exposure models, joint distribution functions of degradation paths, and dependent random effects stochastic processes.

Z. Li · C. Han · D. W. Coit (✉)
Department of Industrial & Systems Engineering, Rutgers University, Piscataway, NJ, USA
e-mail: coit@soe.rutgers.edu

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_19

1 General Degradation Processes

One of the common degradation models is a random effects model, which is a time-dependent function with coefficients as random variables [4, 5]. To be specific, the observed degradation data for unit i at time t_j can be given by Eq. 1:

$$X_{ij}(t_j) = \eta_{ij} + \varepsilon_{ij} = \eta(t_j; \mathbf{\Phi}, \mathbf{\Theta}_i) + \varepsilon_{ij} \tag{1}$$

where $\eta(t_j; \mathbf{\Phi}, \mathbf{\Theta}_i)$ is the actual path of unit i at time t_j, with unknown parameters; $\mathbf{\Phi}$ is the vector of fixed-effect parameters, which is usually common for all units; $\mathbf{\Theta}_i$ is the vector of the i-th unit random-effect parameters, representing individual units; and $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ is the measurement error with constant variance σ_i^2. For a monotonic function, the reliability evaluation can be expressed as Eq. 2:

$$R(t) = \Pr\{X_{ij}(t) < H\} \tag{2}$$

H represents a known failure threshold. Figure 1 shows an example of this type of degradation model, whose degradation function is given by Eq. 3:

$$\eta(t) = \Theta_1 + \Theta_2 t \tag{3}$$

Θ_1 represents the common initial degradation amount of all tested units, which follows normal distribution, $N(10, 3)$; Θ_2 represents the degradation rate, which varies from unit-to-unit, and also follows a normal distribution, i.e., $N(7, 3)$. The degradation path of the given example can be observed in Fig. 1

Random effects models have been used to analyze degradation data of GaAs lasers [6], solid oxide fuel cells [7], modules of fighter aircrafts [8], etc. However, the temporal variability has not been considered into the random effect model. For example, when the time-relevant deterioration function is linear, a single inspection

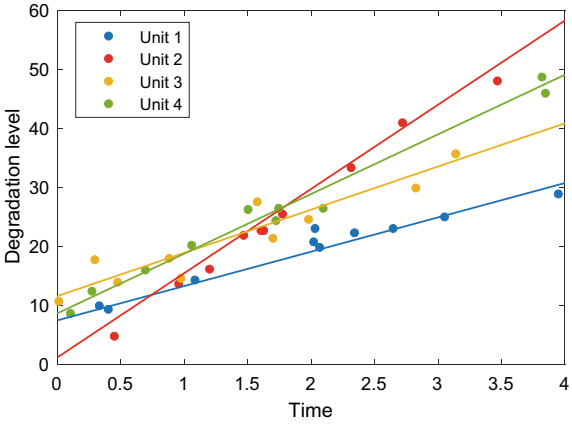


Fig. 1 An example to the random effects degradation model

already completely reveals the future deterioration evolution, just as Fig. 1 shows. Therefore, the random effects model has also been combined or extended with other stochastic process-based degradation models, by letting the physical related parameters in the corresponding degradation model to be a random variable, such as combining with a gamma process [9], inverse Gaussian process [10, 11], Wiener process [12–14], etc.

Another type of degradation model is Markov degradation processes. The Markov processes can reflect the temporal variability of the degradation behavior, by letting the degradation level change at time steps or continuously depending on the previous state. Although the Markov chain can be used to access the condition and reliability of the system [15], for a component experiencing deterioration processes, such as battery degradation, concrete degradation, wear metal corrosion and fatigue, the deterioration process is usually a continuous process which can be difficult to be divided into discrete states. The stochastic processes as gamma process, Gaussian process, compound process, etc., would be more suitable for degradation processes modeling.

The compound Poisson process [16, 17] is a typical representative of the jump-shock degradation model. Compound Poisson process assumes that the degradation process is caused by incoming shocks in finite time intervals, whose interarrival time follows the exponential distribution and the size are *iid*-positive random variables. The total degradation level of one component at time t can be given by Eq. 4:

$$X(t) = \sum_{i=1}^{N(t)} D_i \tag{4}$$

where $N(t)$ is a Poisson random variable, D_i are iid-positive random variables. Figure 2 shows an example of a compound Poisson process. The compound Poisson process has been applied in data analysis to ultra-thin gate oxide data [18], furnace wall degradation [19], etc.

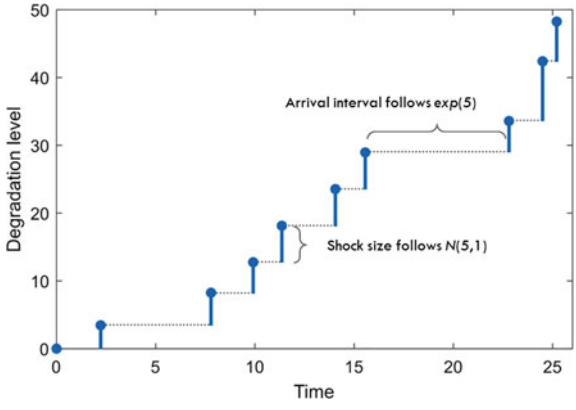


Fig. 2 Example of a compound Poisson process

The Weiner process or Brownian motion with drift is a continuous non-monotonic stochastic process, whose degradation level at time t for a linear process can be given by Eq. 5:

$$X(t) = \mu t + \sigma W(t) \tag{5}$$

where μ is the drift parameter and σ is the volatility parameter, $W(t)$ represents a standard Brownian motion. The Brownian motion with drift has independent increments, for any independent time interval. Therefore, Brownian motion with drift is suitable to describe the degradation process with alternating increases and decreases of degradation resistance. The cumulative distribution function (cdf) of first hitting time $T_D(z)$ when the degradation process first time hits level z can be given with Eq. 6:

$$\begin{aligned} F_{FHT}(t, z) &= \Pr[T_D(z) \leq t] \\ &= \Phi\left(\frac{\mu t - z}{\sigma\sqrt{t}}\right) + \exp\left(\frac{2\mu z}{\sigma^2}\right)\Phi\left(-\frac{\mu t + z}{\sigma\sqrt{t}}\right) \end{aligned} \tag{6}$$

Figure 3 shows an illustrative example of Brownian motion with drift, whose parameters are $\mu = 2$ and $\sigma = 3$. Ye et al. [20] developed a mixed effects model which has several existing Wiener processes as its limiting cases and applied the model to the wear problem of magnetic heads of HDDs, and a light intensity degradation problem of light-emitting diodes. Liao et al. [21] adopted the Brownian motion with drift to model a step-stress accelerated degradation test problem On the basis of Brownian motion, the optimization approach to the common variables, such as sample size, measurement frequency, and termination time optimization were studied. Son et al. [22] combined a principal component analysis approach with Brownian motion and applied the combined model to multivariate time series data that are generated via a thermo-dynamical simulation model for the aircraft engine. The estimation result for the remaining useful life indicates the superiority of their Wiener-based model to the non-probabilistic model with similarity-based prognostic method. Dong and Cui [23] developed a Wiener-gamma degradation model, where the failure thresholds including an alarm threshold, and thresholds regarding degradation amount and duration are used to model system reliability.

However, as a non-monotonic degradation process, Brownian motion with drift may have negative increments and is inadequate in modeling deterioration which is monotone. In contrast, a gamma process and inverse Gaussian process are more suitable for modeling monotonic degradation process. Gamma process is a non-negative stochastic process, which has following properties: (1) the increments $\Delta X(t) = X(t + \Delta t) - X(t)$ follows gamma distribution, and (2) the increments $\Delta X(t)$ in the gamma process are independent.

$$X(t + \Delta t) - X(t) \sim g(x; \alpha(t + \Delta t) - \alpha(t),\ \beta)$$

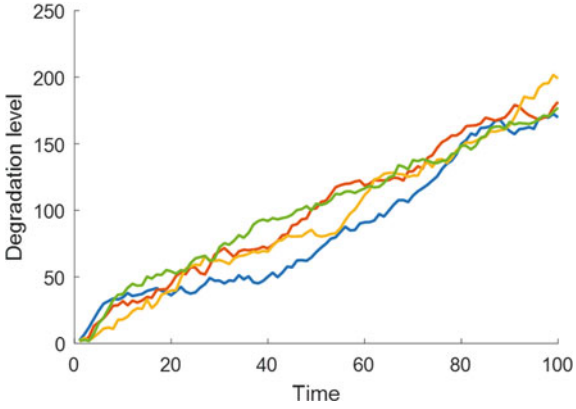


Fig. 3 Example of a Brownian motion with drift

$$g(x; \alpha(t+\Delta t)-\alpha(t),\ \beta) = \frac{\beta^{\alpha(t+\Delta t)-\alpha(t)} x^{\alpha(t+\Delta t)-\alpha(t)} \exp(-\beta t)}{\Gamma(\alpha(t+\Delta t)-\alpha(t))} \tag{7}$$

Figure 4 shows an example of a non-linear gamma process whose parameters $\alpha(t) = 6e^{0.2t} - 6$, and $\beta = 200$. The gamma process can be regarded as having an infinite number of jumps in finite time intervals, which is suitable for describing gradual damage by continuous use.

There is also significant amounts of degradation research based on gamma process models. Park and Padgett [24] provided the approximate distribution of the first hitting time, which is given by a form of the Birnbaum-Saunders distribution, given the initial degradation level x_0. Tseng et al. [25] introduced a step-stress accelerated degradation test model when the degradation process follows a gamma process

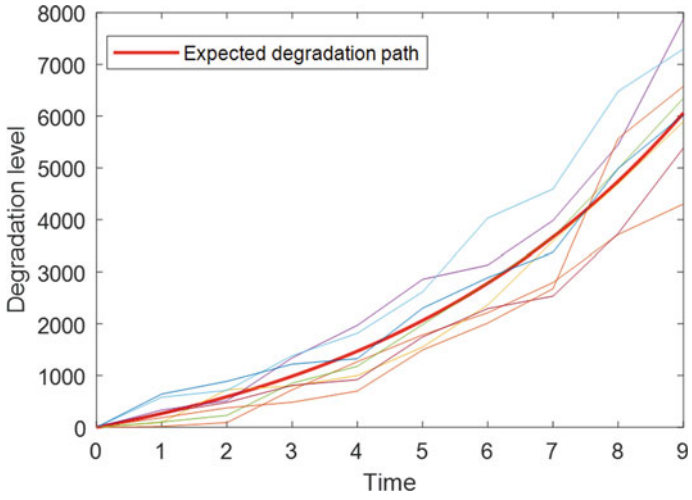


Fig. 4 Example of a non-linear gamma process

and determined the corresponding optimal settings such as sample size, measurement frequency, and termination time. Pan and Balakrishnan [26] discussed the modeling for multiple performance characteristics based on gamma processes, by using a bivariate Birnbaum-Saunders distribution and its marginal distributions to approximate the reliability function. Tsai et al. [27] used a two-variable accelerated degradation test model based on a gamma process to fit the lumen degradation of LEDs. Zhou et al. [28] introduced a two-stage degradation model which contains a time-to-event initiation stage and a gamma process propagation stage, and applied it to the rebar corrosion modeling. Li et al. [29] used a two-stage degradation model, where both degradation stages are modeled by gamma processes, to characterize the rebar corrosion process. Then, they used machine learning approaches to control the prediction bias of the gamma-gamma two-stage model.

The inverse Gaussian process is also a monotonic stochastic process with independent increments, for any independent time interval. For this model, the degradation increment follows inverse Gaussian distribution:

$$X(t+\Delta t) - X(t) \sim IG(x;\, \Lambda(t+\Delta t) - \Lambda(t), \eta[\Lambda(t+\Delta t) - \Lambda(t)]^2)$$

where $\Lambda(t)$ is a monotone increasing function.

$$f_{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}} \exp\left[-\frac{b(x-a)^3}{2a^2 x}\right] \tag{8}$$

and $x > 0$. Therefore, an inverse Gaussian process is strictly monotone. Peng et al. [30] inferred the cdf of first hitting time of an inverse Gaussian process in terms of degradation level. The same author also demonstrated the applicability of the Bayesian method for degradation analysis with the inverse Gaussian process models. Ye et al. [11] studied the optimal constant-stress accelerated degradation tests planning when the underlying degradation follows the inverse Gaussian process. Guo et al. [31] proposed an improved inverse Gaussian process which considers the dependency between degradation increments and prior degradation states and performed reliability analysis to the crack length growth data based on the proposed model.

1.1 Dependency in Reliability Models

Functional or physical dependence refers to the phenomenon that failure or degradation of one component directly influences others, perhaps in a cause-and-effect or symbiotic relationship. For example, the degradation of the reinforcement concrete used for road bridges and the embedded rebar affects and promotes each other. Specifically, the degradation of concrete leads to the appearance of cracks, which allows moisture and ions to erode the embedded rebars. Meanwhile, the expansion of rebars during the corrosion process in turn promotes the propagation of concrete

cracks. Therefore, the degradation processes of concrete and rebar in the same block of reinforcement concrete are dependent. According to Nicolai and Dekker [32], stochastic dependence (*s*-dependence) refers to the dependence between components that is expressed by a change of the conditional distribution of any of the related components when the other components are altered.

Covariance or correlation are measures of how much two random variables vary together, shown as Eqs. 9 and 10. When covariance is greater than 0, random variables at time t, $X_1(t)$ and $X_2(t)$, appears to be dependent.

$$Cov(X_1(t), X_2(t)) = E[X_1(t)X_2(t)] - E(X_1(t))E(X_2(t)) \tag{9}$$

$$\rho_{X_1(t)X_2(t)} = \frac{Cov(X_1(t), X_2(t))}{\sigma_1\sigma_2} \tag{10}$$

Common cause failure [33] is one of the most widely studied dependent failure modes in terms of failure time. It often occurs when multiple components fail due to an external shock or event that simultaneously causes multiple failures. The frequency of common-cause events leads to dependent failures.

A load-sharing system refers to a system consisting of multiple components, where when any component fails, the same workload must be shared by the remaining surviving components, resulting in an increased load on each surviving component. Ye et al. [34] derived the time to degradation failure of such the system, estimated the system probability of failure, and investigated the optimal designs to minimize the long run average cost of a future system.

To model the dependent failure times, Markov chains [35, 36] and joint distribution functions (include copula function) [37–39] and other methods have often been adopted.

2 Markov Chain in Multistate Degrading System

It is common that a variety of systems experience deterioration and unforeseen shock damages. A complicated system with many components degrades over time as its constituent parts fail and deteriorate. Markov chains are an analytical tool that can be used to model state degradation behavior for the components as they collectively degrade within a system.

The majority of repairable systems, including computers, power generators, nuclear systems, and airplanes, can be brought back to operational status through repair or refurbishment. In contrast, non-repairable systems are replaced wholly when they fail. Recent technological advancements have led to hybrid systems with a variety of functions that are more complicated and hybrid in nature. Each subsystem and component in these complex systems may be subject to one or more degradation processes. Assessing the reliability and planning the maintenance of

such complex systems requires a more accurate evaluation of the degradation status of its components and the propagation of degradation processes in the future [40]. Existing literature on maintenance optimization gives a number of methodologies for reliability analysis of such complex systems [41, 42]

Reference [43] categorized the interdependencies between components as economical, structural, and stochastic (*s*-dependence). As a result of economic dependence, maintenance costs can be decreased by performing maintenance on many components simultaneously rather than individually. Structural dependence is founded on the premise that specific components fundamentally compose a unit (subset of the system), and hence maintenance of one component in a unit necessitates the maintenance of other components. *s*-dependence refers to system interaction in which one component's degradation or failure process is influenced by the degradation status of one or more interrelated and/or adjacent system components [44, 45]. Several research efforts have addressed *s*-dependence between components with either two possible degradation states (binary states) or multiple discrete degradation stages [46, 47] or [48–51]. Not as much progress has been made in examining *s*-dependence among multistate components with more than two but a finite number of alternative degradation stages for each component [52, 53].

Numerous complex systems, such as aircraft engines, wind turbines, and power generation systems, are composed of multiple components arranged serially to complete successive missions [54, 55]. During the interim between two missions, maintenance procedures can be performed to increase the likelihood of successfully completing the subsequent mission. Due to the limited availability of maintenance resources, it is crucial to choose the maintenance strategy (e.g., the subset of components to maintain and the number of maintenance operations) in light of the system requirements. This sort of maintenance has received increasing attention more recently [55]. Many research studies on maintenance policy have recommended formulating it as an optimization problem, assuming that all components are stochastically independent. However, often components in several complex engineering systems, particularly mechanical and electrical systems, do affect the degrading behavior of other components.

The degradation of a component in an electronic system may result in a considerable rise in temperature, which in turn may accelerate the degradation of other temperature-sensitive components [56, 57], which may or may not be directly related. Similarly, in power grid systems, the deterioration of generators or transformers in a sub-network may raise the demand on other network elements, creating an increase in their loading profiles and so accelerating their deterioration [48].

A system could be defined as a multistate system if there is a finite number of discrete performance rates. One of the approaches to modeling the multi-state system is using the Markov chains to analyze the system state transition process [58]. The system states are defined by the combination of component states. As the number of components increases, the state of system could increase rapidly. Therefore, the system structure function should be adopted to help define system states [59]. Figure 5 shows the state transition behavior of dependent Markov Chains at a certain time, which consists of N multistate dependent components and each

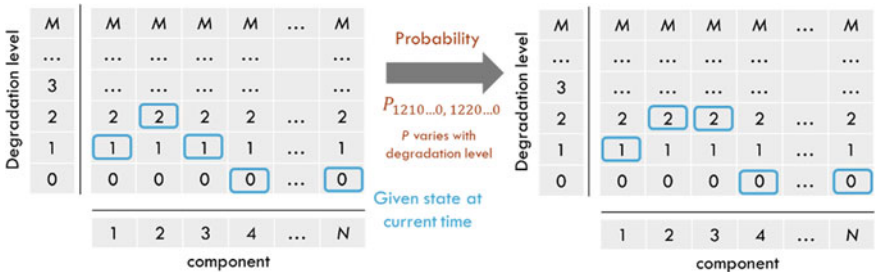


Fig. 5 State space of N dependent Markov chains at a certain time

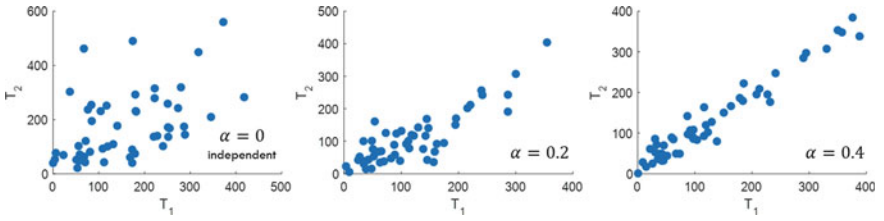


Fig. 6 Failure time for two dependent components modeled by Markov chain with different level of dependency

component has $M + 1$ different performance levels. Figure 6 shows the simulated failure time of a dependent Markov chain. In the simulation, there are 2 different components. For each component, there are states 0 to 5, state 5 is failure state. Transition time from state 0 to 1 for components 1 and 2 follow, $T_{01} \sim \exp(1/80)$ and $T_{02} \sim \exp(1/100)$. While degrading, the transaction rate λ increases by $\lambda(1+\alpha)^i$, i is the higher degradation state between two components. Degradation for one component can then affect λ for another component as well.

Xue and Kang [59] combined Markov processes with *s*-coherent multistate system theory, performed reliability analyses to the multistate system. They also conducted the general reliability function for *s*-coherent multistate systems. Refs. [60–62] discussed the optimal maintenance strategy for multistate systems which have discrete degradation states resulting from cumulative damage. Zhang et al. [63] assumed that the system deterioration could be stochastic and the maintenance of such a system could be "imperfect", which indicates that the component after repair is not as good as new. They studied the replacement policy based on the failure number of such a multistate system. The same researchers [64] also extended their work by considering component working age and failure times in an imperfect maintenance policy. Since imperfect maintenance may not necessarily result in a perfect condition, the maintenance cost is less than if a perfect action were performed. However, the difficulty of the selective maintenance optimization problem is increased by including imperfect maintenance activities. Moreover, the majority of prior research

in hybrid maintenance policy has assumed that the quality of incomplete maintenance activities is predetermined [52, 65].

In a complex and hybrid system, components might be viewed as inter-dependent; hence, the component dependency must be considered when modelling system degradation and planning maintenance measures [66, 67]. Liu and Huang [68] proposed an optimal replacement strategy for multistate systems incorporating imperfect maintenance quality. They assumed that the transition rate in the Markov model of multistate system to be "age correlated". That is, the component would be restored to its best functional state, but with higher state transition intensities after replacement, which means that components degrade more rapidly to a lower performing state. A non-homogeneous continuous time Markov model, where the state transition intensity varies with time, was applied to model the aging of components in such a system. Shahraki et al. [69] considered *s*-dependency between correlated components. In their model, component degradation is assumed to follow a continuous time Markov process, with transition times between component states following an exponential distribution. The degradation rate of a component consists of two elements: the intrinsic degradation rate, and the interaction effect on degradation rate caused by other degrading components. The interaction is a function of overall system performance at time t, as well as number of influencing components transitioning to a lower state at time t. By considering interaction effects in the state transition rate, they addressed the dependency between components in their Markov model, and the selective maintenance optimization problem could be determined.

3 Shared Shock Exposure Models

The single degradation mechanism has been extensively applied to numerous items. Due to the complexity of the product and the unpredictability of external circumstances, the degradation process is frequently accompanied by a range of failure mechanisms in practice. Moreover, failure mechanisms frequently exhibit coupling relationships that influence the degrading characteristics of the product. This differs from the conventional single-mechanism degradation processes. For electronics and mechanics, competing failure processes have been widely studied. In this chapter, multiple dependent competing failure processes are addressed and further analyzed as an example.

There are numerous system design and optimization problems where each component is simultaneously exposed to deterioration and shocks, leading to *s*-correlated or *s*-dependent component failure processes. When there are two or more dependent failure processes and several components, many standard reliability techniques can be insufficient or inappropriate. In reliability modeling, there are difficult problems caused by the inter-dependency between failure processes and between component failure times. Shared shock exposure models create such a scenario where a system

is exposed to random shocks, and when the system is exposed to a shock, all components experience shock effects. Degradation is the sum of a continuous degradation process and cumulative effect of shocks.

Before introducing the shared shock exposure model, it is necessary to briefly introduce the random shock model [70–73], which describes the degradation model caused by the impact of environment such as sudden and unexpected usage loads, e.g., accidental dropping onto hard surfaces. Random shock models can be classed into four categories: (1) extreme shock model; (2) cumulative shock model; (3) run shock model; (4) δ-shock model. Alternatively, the types of failures due to shocks can be divided into soft and hard failures. Soft failure occurs when the system overall degradation level, which is contributed by an internal degradation process and additional abrupt shock damage, exceeds a preset critical threshold H. Hard failures occurs when the load magnitude from a single shock exceeds a critical strength level D. Peng et al. [74] studied the system considering both soft and hard failure as a dependent competing failure process, by assuming the same random shock can not only contribute to soft failure, but can also trigger hard failure, depending on which failure occurs first. In their model, the random shock events occur as a Poisson process with arrival rate λ; shock damage size are *i.i.d.* random variables following a certain distribution. The probability that the component survives the stress from the i th shock is:

$$P(W_i < D) = F_W(D), \text{for } i = 1, 2, \ldots, \infty \tag{11}$$

The cumulative damage caused by shocks until time t can be given as:

$$S(t) = \begin{cases} \sum_{i=1}^{N(t)} Y_i, & \text{if } N(t) > 0 \\ 0, & \text{if } N(t) = 0 \end{cases} \tag{12}$$

where $N(t)$ is the total number of shocks that have arrived by time t. Then the cdf of total degradation at time t is:

$$F_X(x, t) = \sum_{i=0}^{\infty} P(X(t) + S(t) < x | N(t) = i)\ P(N(t) = i) \tag{13}$$

Figure 7 shows a soft failure process and a hard failure process.

Song et al. [75] extended the above failure model into a system with multiple dependent component failure processes. In their proposed model, the same random shock contributes to both soft failure and hard failure for all components in the system, which results in s-dependent component failure processes and failure times. If a component fails relatively frequently, then the number of underlying shocks are more likely to be relatively large. Since these shocks affect all components, they may also cause other components to fail more frequently. Consider n components in such a system with series configuration, the system reliability function can be given as Eq. 14:

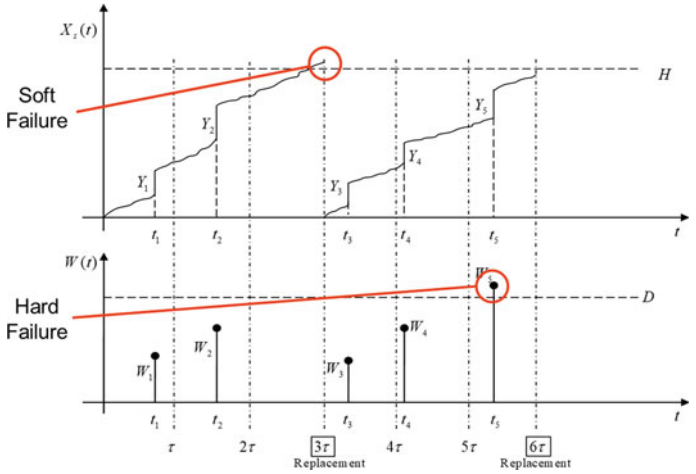


Fig. 7 Two dependent competing failure processes: soft failure process and hard failure process

$$R(t) = P\{[W_{11} < D_1, W_{12} < D_1, \ldots, W_{1N(t)} < D_1, X_{s1}(t) < H_1] \cap [W_{21} < D_1, W_{22} < D_1, \ldots, W_{2N(t)} < D_1, X_{s2}(t) < H_2] \cap \ldots \cap [W_{n1} < D_1, W_{n2} < D_1, \ldots, W_{nN(t)} < D_1, X_{sn}(t) < H_1]\} \quad (14)$$

where W_{ij} is the j th shock for component i; $X_{si}(t)$ is the cumulative degradation level for component i at time t.

They proved such dependency by demonstrating that the covariance of these failure events for any two components is positive, given the presence of shocks. Song et al. [76] then conducted a more advanced model considered the scenario where shocks with specific scale, frequency or function that can selectively affect one or more components in the system but not necessarily all components. The authors used a MEMS (micro-electromechanical systems) oscillator as a typical system to illustrate the reliability analysis and maintenance policy based on the proposed model. Furthermore, the same researchers studied the situation [77] where two failure processes for each component are dependent in additional ways other than external shocks. In their consideration, system shocks are transmitted to the component level, but likely to be dependent. If damage to one component is large, it is probabilistically likely damages on other components are also large. Yousefi et al. [78] studied the maintenance optimization model for such a shared shock exposure model by minimizing the average long run maintenance cost rate.

4 Joint Function of Degradation Paths

The modeling of a degrading process is usually based on two modeling frameworks, the general path model and the stochastic process model. The general path model is established by referring to some form of a regression model, and by defining the parameters as random variables to account for the variability. Lu et al. came up with a model for general path process and joint distribution function [79]. The equation for the proposed model is shown as Eq. 15:

$$y_{ijk} = g_{ij}\left(t_{ik}, \boldsymbol{x}_i, w_{ij}; \boldsymbol{\eta}_j\right) + \epsilon_{ijk} \tag{15}$$

y_{ijk} represents the measurement for unit i on the jth degradation measure at time t_{ik}. $\boldsymbol{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)'$ denotes a vector that contains all the covariate variables for unit i, where x_{iv} denotes the vth covariate for unit i. $\epsilon_{ijk} \sim N\left(0, \sigma_j^2\right)$ is the independent measurement error which has a common variance of σ_j^2 for measurements on the jth degradation character. w_{ij} stands for the random effects parameter associated with unit i for the j degradation characters, which follows a multivariate normal distribution (MVN) given by $\boldsymbol{w}_i = \left(w_{i1}, \ldots, w_{ij}\right)' \sim MVN(0, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma}$ is a variance–covariance matrix. $g_{ij}(\cdot)$ can be highly non-linear and based on a physical model. The dependent paths are caused by the joint distribution function for $\boldsymbol{w}_i$. Additionally, Ref. [79] also derived the reliability analysis for the general path model and a parameter estimation approach based on an expected maximum algorithm. The reliability assessment for the general path model can be expressed by Eq. 16 below:

$$\begin{aligned} R(t) &= \Pr\{T_1 \geq t, \ldots, T_J \geq t\} \\ &= Pr\left\{g_1\left(t, \boldsymbol{x}, w_1; \boldsymbol{\eta_1}\right) \geq d_1, \ldots, g_J\left(t, \boldsymbol{x}, w_J; \boldsymbol{\eta_J}\right) \geq d_J\right\} \\ &= Pr\left\{\ln\left[-g_1\left(t, \boldsymbol{x}, w_1; \boldsymbol{\eta}_1\right)\right]\right. \\ &\left.\leq \ln(-d_1), \ldots, \ln\left[-g_J\left(t, \boldsymbol{x}, w_J; \boldsymbol{\eta}_J\right)\right] \leq \ln(-d_J)\right\}. \end{aligned} \tag{16}$$

The reliability $R(t)$ depends on the joint failure time distribution for all the J degradation character measurements. It also depends on the functional form of $g_j(\cdot)$, which is based on a physical model.

Ye et al. [80] proposed a multi-stress acceleration model with interaction effects based on the general path model. In their model, the effects of the multiple stresses applied on the component can be divided into main stress effects and interaction effects. The degradation rate $\mu(\boldsymbol{S})$ under N multiple stresses $\boldsymbol{S}$ can be expressed as Eq. 17:

$$\mu(\boldsymbol{S}) = \prod_{k=1}^{N} \alpha_k \, h_k(S_k, \beta_k) \tag{17}$$

$h_k(\cdot)$ denotes the explicit functional form; S_k is the k-th stress; β_k is a constant. α_k is the interaction effect term, which can be expressed as:

$$\alpha_k = \lambda_k \cdot g_k((\boldsymbol{S}\backslash\{S_k\}), \gamma_k) \tag{18}$$

where λ_k and γ_k are unknown parameters; $\boldsymbol{S}\backslash\{S_k\}$ denotes the stress variables except S_k; $g_k(\cdot)$ is the explicit functional from of the interaction, which can be obtained based on the exponential and power laws [81, 82].

The other method for dependent degradation modeling is based on stochastic processes. Compared with the general path model, this type of method focuses on establishing a joint distribution function of the multivariate degradation process. Mercier and Pham [83] adopted a bivariate non-decreasing Lévy process to model a two-unit system, and presented the failure threshold under various cases. Trivariate reduction approach, proposed by Cherian [84], provides an approach to construct the bivariate gamma process. The framework of the approach can be summarized as following: let $\{Y_1(t)\}_{t\geq 0}$, $\{Y_2(t)\}_{t\geq 0}$, and $\{Y_u(t)\}_{t\geq 0}$ be three independent univariate gamma process with parameters (α_1, β), (α_2, β), and (α_u, β). Let $X_1(t) = Y_1(t) + Y_u(t)$, and $X_2(t) = Y_2(t) + Y_u(t)$. Therefore, the process $\{X(t)\}_{t\geq 0} = \{(X_1(t), X_2(t))\}_{t\geq 0}$ is a bivariate subordinator with gamma marginal processes and parameters (a_i, β), where $a_i = \alpha_i + \alpha_u$, $i = 1, 2$. The linear correlation coefficient between two degradation processes $X_1(t)$ and $X_2(t)$ is $\rho = \frac{\alpha_u}{\sqrt{a_1 a_2}}$.

Liu et al. [85] derived the joint probability distribution function (pdf) and cdf of such a bivariate gamma process, and investigated the condition-based maintenance policy for a two-unit system in the finite-horizon setting. Pan and Balakrishnan [26] developed a bivariate gamma process by considering a constant correlation coefficient of the increment between two gamma processes, and approximated the reliability by a bivariate Birnbaum–Saunders distribution and its marginal distributions. Liu et al. [86] and Dong et al. [87] used the covariance matrix to capture the dependency between two degradation processes, and developed the degradation model based on a Wiener process.

In addition to these approaches, the copula function is another frequently used approach for constructing joint distribution functions. Copula functions are functions that connect multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copulas are multivariate distribution functions whose one-dimensional intervals are uniform on the (0,1) interval. For mathematical definitions and properties of copula, please refer to [88]. In general, the framework of copula can be written as Eq. 19:

$$C(u_1, u_2, \ldots, u_d) = P(U_1 < u_1, U_2 < u_2, \ldots, U_d < u_d) \tag{19}$$

where U_k is a uniform random variable between [0, 1]. The relationship between multivariate distribution functions and their univariate margins in terms of copula can be explained by the Sklar's theorem [88]:

Let H be a joint distribution function with margins $F_1, F_2, \ldots, F_d$. Then there exists a copula C such that for all $x_1, x_2, \ldots, x_d$ in $\overline{\boldsymbol{R}}$,

$$H(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots F_d(x_d)$$

If F_k are continuous, then C is unique. Conversely, if C is a copula and F_k are distribution functions, then the function H is a joint distribution function with margins $F_1, F_2, \ldots, F_k$.

By letting F_k characterize the distribution of one of the dependent degradation processes, with a proper copula function, the joint distribution function of the dependent degradation processes can be obtained. Therefore, the corresponding pdf can be derived as Eq. 20:

$$f(x_1, x_2, \ldots, x_d) = c(F(x_1), F(x_2), \ldots, F(x_d)) \prod_{k=1}^{d} f_k(x_k) \tag{20}$$

where $f_k(x_k)$ is the marginal pdf of X_k and $c(F(x_1), F(x_2), \ldots, F(x_d))$ is the copula density function. Common copula functions include Gaussian copula [89, 90], Clayton copula [91, 92], Gumbel copula [90], and Frank copula. Figure 8 shows common copulas and their applicable scenarios for a particular system reliability model. Figure 9 shows Gumbel copula and Clayton copula, which have strong head correlation and tail correlation, respectively.

Fang et al. [94] developed a framework for bivariate dependent degradation model, where the dependency was charactered by a copula function. The parameter estimation approaches were also derived. Liu et al. [95] used a Clayton copula function to construct a bivariate gamma process, as well as, conducted the reliability analysis and life cycle cost analysis. Xu et al. [51] applied a vine copula to reliability functions associated with three performance characteristics, while each performance characteristic had its own failure threshold. The reliability joint distribution of all degradation metrics of a smart electricity meter was obtained. Li et al. [93] who solved an optimization problem of the multi-dimensional gamma process model based on copula to determine inspection times and preventive maintenance policies. Sun et al. [13] analyzed the bivariate and multivariate dependent accelerated degradation data respectively by means of the copula theory.

5 Dependent Random Effects Stochastic Process

In addition to constructing the joint distribution function of the degradation level of the stochastic process directly, it is also a practical modeling approach to examine the correlation of the stress factors, which reflects the parameters in the stochastic processes, to construct the correlation model at the parameter level of the stochastic degradation process. For the dependent random effects stochastic process model, the

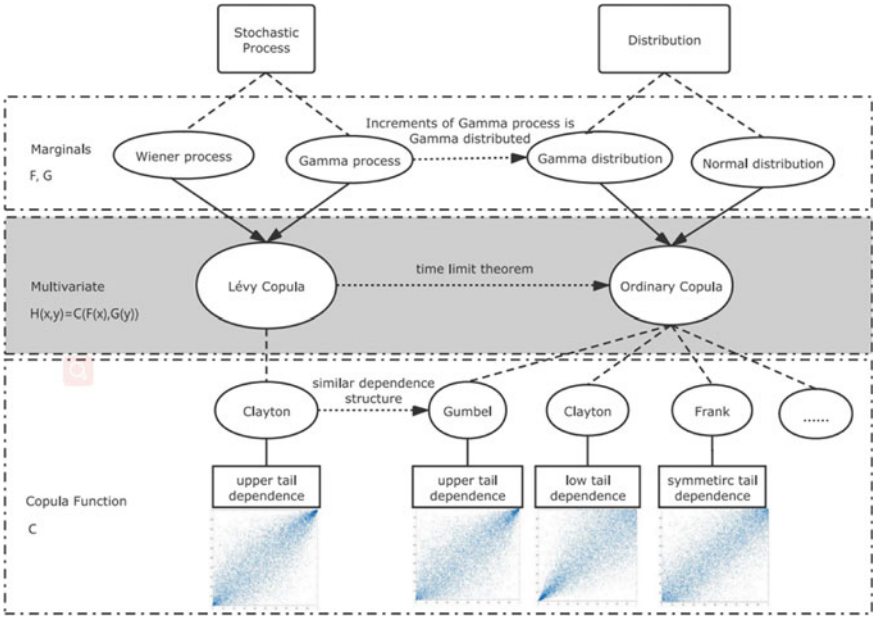


Fig. 8 Overview diagram of copula method [93]

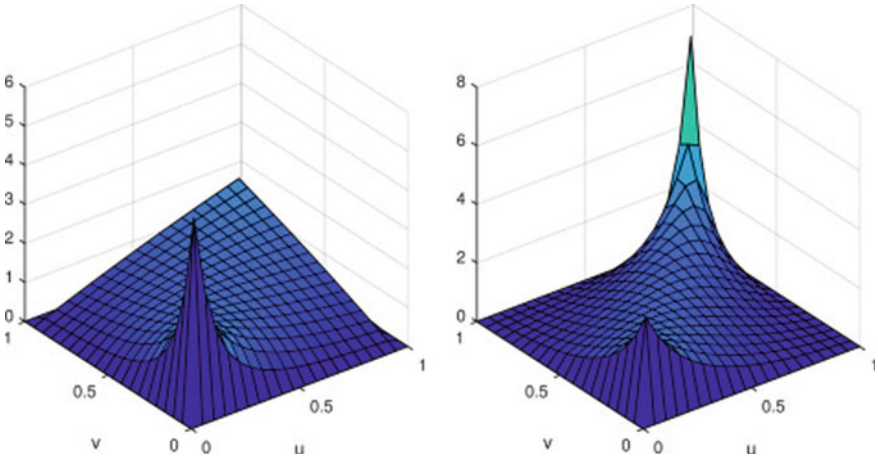


Fig. 9 Graphs of Clayton and Gumbel copulas [95]

degradation processes are in accordance with a stochastic progress and the process parameter(s) are random variables, which could share a random variable in a function, or can be described by a joint distribution function, such as copula functions.

Cui and Song [96] developed a bivariate gamma degradation model, where the dependency between the two degradation processes is captured by a common random effect naturally. The equation of two correlated gamma process can be given as follows:

$$X_1(t)|\lambda \sim G(\lambda\mu_1, a_1\eta(t; b_1)),\ X_2(t)|\lambda \sim G(\lambda\mu_2, a_2\eta(t; b_2))$$

where $\lambda \sim G(\beta, \beta)$. $\eta(t; b_k)$ is a non-decreasing function of t with $\eta(0; b_k) = 0$, and the parameter b_k has a critical role in transforming the time scale. They also further developed the MLEs of the model parameters and the RUL prediction methods. Xu et al. [97] proposed a bivariate Wiener process model, where the degradation process of two performance characteristics share a common random coefficient parameter. To be specific, the degradation process of the sth performance characteristic is modeled as Eq. 21:

$$Y_s(t) = \alpha\beta_s h_s(t, \gamma_s) + \sigma_s B_s(h_s(t, \gamma_s)), \quad s = 1, 2 \tag{21}$$

where β_s and σ_s denote the drift parameter and diffusion parameter. $h_s(t, \gamma_s)$ is a nondecreasing function and $h_s(0, \gamma_s) = 0$. $B_s(\cdot)$ is a standard Brownian motion. The dependency between two performance characteristics is determined by parameter α, which follows a normal distribution $N\left(1, \sigma^2\right)$. The bivariate distribution of $Y_1(t)$ and $Y_2(t)$ can be derived based on the distribution of α, which is shared by both performance characteristics. Hong et al. [98] proposed a dependent Wiener process where the parameter of the dependent Wiener process can be described by a common distribution, which is constructed via copula function. The model can be expressed as Eq. 22:

$$D_j(t) = D_j\left(t; \mathbf{x}, \omega_j\right) = \omega_j \Lambda_j(t; \mathbf{x}) + \sigma_j B\left(\Lambda_j(t; \mathbf{x})\right) \tag{22}$$

$\boldsymbol{\omega} = \left(\omega_1, \ldots, \omega_p\right)' \sim C\left(F_1(\omega_1), \ldots, F_P\left(\omega_p\right)\right)$; $C(\cdot)$ is a selected copula function, $F_j\left(\omega_j\right)$ is the marginal cdf for parameter ω_j. Conditional on ω_j, each $D_j(t)$ is an independent process.

Consider a scenario in which components exist in a shared environment, and factors like temperature, humidity, etc. can affect degradation of all or some components. Therefore, components in the same clusters degrade together or dependently. For example, drive combs in MEMS accelerometers experience the similar harsh environment. Figure 10 shows an illustrative example of two clusters of degradation paths for seven identical components in a system. Yousefi et al. [99] proposed a degradation model considering clusters of components, where degradation is based on a random effects gamma process. They further developed reliability analysis for series and parallel layout systems. To be specific, each degradation path is modeled as a random effects gamma process as showing below:

$$X_i(t_2) - X_i(t_1) \sim Ga(x_i; v_i(t_2) - v_i(t_1), \theta(i))$$

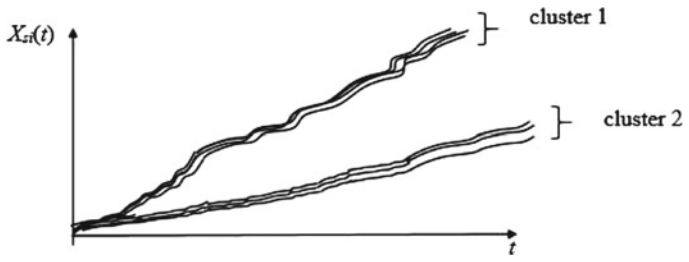


Fig. 10 Two clusters of degradation paths for seven identical components in a system [99]

where $\theta(i) = \alpha_{0,i} + \alpha_{1,i}\theta_1 + \cdots + \alpha_{k,i}\theta_k$ is the shape parameter; where each θ_j corresponds to a cluster; $\alpha_{j,i}$ is a sensitivity factor which defines the degree which component i belongs to cluster j; $v_i(t)$ is the scale parameter.

Wu et al. [100] proposed a multivariable degradation model based on a Wiener process and considered the dependency in the same clusters. The authors utilized the expectation–maximization (EM) algorithm to estimate the model parameters and used a tangent approximation approach to conduct the reliability function of the proposed model.

6 Conclusion

In this chapter, we first introduced general degradation models, including random effects model, compound Poisson process, Wiener process, gamma process, inverse Gaussian process. Among them, the general degradation path model is based on some form of a regression model, and the compound Poisson model is suitable to describe the system degradation caused by finite shocks. The remaining processes are suitable to describe the gradual degradation process of the system, where the Wiener process is non-monotonic and the others are monotonic degradation processes.

By investigating and identifying the multiple degradation states, the dependent Markov chain is an intuitive approach to modeling the dependent degradation processes. By clarifying the state space of each component in detail, we can use Markov chains to model the obtained multi-state system. The dependence between different components is mainly reflected in the state transition probability of the Markov chain. This transition probability can by influenced by age of the components, the number of failures, or the current state of other components.

Shared shock exposure models assume that the system is exposed to random shocks, and when the system is exposed to a shock, all components experience shock effects. Degradation is the sum of a continuous degradation process and the cumulative effect of shocks. The shock may affect all the components or only some of the components in the system. Also, the dependency between components not only can come from the impact of shocks, but the inherent dependency of components themselves on each other.

Apart from the dependent Markov Chain and shared shock exposure model, a common idea to address system condition in terms of multiple dependent degradation processes is to construct a joint distribution function of the marginal degradation processes. By introducing coefficients that follow a certain multivariate distribution, a general path model can be applied to characterize such dependencies. A covariance matrix can also contribute to constricting multivariable Wiener process. For multivariable gamma processes, tri-variate reduction could be a reasonable approach in assisting the modeling. Alternatively, copula functions, despite difficulties in selecting a proper copula function or deriving the analytical form for the reliability indices, is an important tool to analyze the multivariate dependent degradation data.

The dependent random effects stochastic process model is dedicated to characterizing the dependency among random processes, by introducing correlated parameters, which can originate from a multivariate distribution or combined and approximated by copula functions. Wiener process model and gamma process model with dependent parameter vector are the examples of this approach. Meanwhile, this chapter introduces degradation process considering clusters of components, where components could suffer from similar environmental stress or impact by inherent bindings and show similar degradation behavior, as an approach to character the dependency of different degradation processes.

References

1. Sari JK, Newby MJ, Brombacher AC, Tang LC (2009) Bivariate constant stress degradation model: LED lighting system reliability estimation with two-stage modelling. Qual Reliab Eng Int 25(8):1067–1084
2. Mi J, Li Y-F, Peng W, Huang H-Z (2018) Reliability analysis of complex multi-state system with common cause failure based on evidential networks. Reliab Eng Syst Saf 174:71–81
3. Rodríguez-Picón LA, Rodríguez-Picón AP, Alvarado-Iniesta A (2019) Degradation modeling of 2 fatigue-crack growth characteristics based on inverse Gaussian processes: a case study. Appl Stoch Model Bus Ind 35(3):504–521
4. Lu CJ, Meeker WQ (1993) Using degradation measures to estimate a time-to-failure distribution. Technometrics 35(2):161–174
5. Zuo MJ, Renyan J, Yam RCM (1999) Approaches for reliability modeling of continuous-state devices. IEEE Trans Reliab 48(1):9–18
6. Duan F, Wang G (2018) exponential-dispersion degradation process models with random effects and covariates. IEEE Trans Reliab 67(3):1128–1142
7. Guida M, Postiglione F, Pulcini G (2015) A random-effects model for long-term degradation analysis of solid oxide fuel cells. Reliab Eng Syst Saf 140:88–98
8. Sohn SY, Yoon KB, Chang IS (2006) Random effects model for the reliability management of modules of a fighter aircraft. Reliab Eng Syst Saf 91(4):433–437
9. Tsai CC, Tseng ST, Balakrishnan N (2012) Optimal design for degradation tests based on gamma processes with random effects. IEEE Trans Reliab 61(2):604–613
10. Hao SH, Jun Y, Berenguer C (2019) Degradation analysis based on an extended inverse Gaussian process model with skew-normal random effects and measurement errors. Reliab Eng Syst Saf 189:261–270
11. Ye ZS, Chen LP, Tang LC, Xie M (2014) Accelerated degradation test planning using the inverse Gaussian process. IEEE Trans Reliab 63(3):750–763

12. Wang XF, Wang BX, Jiang PH, Hong YL (2020) Accurate reliability inference based on Wiener process with random effects for degradation data. Reliab Eng Syst Saf **193**
13. Sun FQ, Fu FY, Liao HT, Xu D (2020) Analysis of multivariate dependent accelerated degradation data using a random-effect general Wiener process and D-vine Copula. Reliab Eng Syst Saf **204**
14. Ye ZS, Chen N, Shen Y (2015) A new class of Wiener process models for degradation analysis. Reliab Eng Syst Saf 139:58–67
15. Siu N (1994) Risk assessment for dynamic-systems—an overview. Reliab Eng Syst Saf 43(1):43–73
16. Sahinoglu M (1992) Compound-Poisson software reliability model. IEEE Trans Softw Eng 18(7):624–630
17. Barbour AD, Chryssaphinou O, Roos M (1996) Compound poisson approximation in systems reliability. Naval Res Logistics (NRL) 43(2):251–264
18. Hsieh MH, Jeng SL (2007) Accelerated discrete degradation models for leakage current of ultra-thin gate oxides. IEEE Trans Reliab 56(3):369–380
19. Zhang J-X, Hu C-H, He X, Si X-S, Liu Y, Zhou D-H (2017) Lifetime prognostics for deteriorating systems with time-varying random jumps. Reliab Eng Syst Saf 167:338–350
20. Ye ZS, Wang Y, Tsui KL, Pecht M (2013) Degradation data analysis using wiener processes with measurement errors. IEEE Trans Reliab 62(4):772–780
21. Liao CM, Tseng ST (2006) Optimal design for step-stress accelerated degradation tests. IEEE Trans Reliab 55(1):59–66
22. Son KL, Fouladirad M, Barros A, Levrat E, Lung B (2013) Remaining useful life estimation based on stochastic deterioration models: a comparative study. Reliab Eng Syst Saf 112:165–175
23. Dong Q, Cui L (2019) A study on stochastic degradation process models under different types of failure Thresholds. Reliab Eng Syst Saf 181:202–212
24. Park C, Padgett WJ (2005) Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. Lifetime Data Anal 11(4):511–527
25. Tseng ST, Balakrishnan N, Tsai CC (2009) Optimal step-stress accelerated degradation test plan for gamma degradation processes. IEEE Trans Reliab 58(4):611–618
26. Pan ZQ, Balakrishnan N (2011) Reliability modeling of degradation of products with multiple performance characteristics based on gamma processes. Reliab Eng Syst Saf 96(8):949–957
27. Tsai TR, Sung WY, Lio YL, Chang SI, Lu JC (2016) Optimal two-variable accelerated degradation test plan for gamma degradation processes. IEEE Trans Reliab 65(1):459–468
28. Zhou J, Coit DW, Nassif H, Li Z (2022) Two-stage degradation modeling combined with machine learning for steel rebar degradation prediction. In: 2022 annual reliability and maintainability symposium (RAMS)
29. Li Z, Zhou J, Nassif H, Coit D, Bae J (2023) Fusing physics-inferred information from stochastic model with machine learning approaches for degradation prediction. Reliab Eng Syst Saf 232:109078
30. Peng W, Li Y-F, Yang Y-J, Huang H-Z, Zuo MJ (2014) Inverse Gaussian process models for degradation analysis: a Bayesian perspective. Reliab Eng Syst Saf 130:175–189
31. Guo JB, Wang CX, Cabrera J, Elsayed EA (2018) Improved inverse Gaussian process and bootstrap: degradation and reliability metrics. Reliab Eng Syst Saf 178:269–277
32. Nicolai RP, Dekker R (2008) Optimal maintenance of multi-component systems: a review, in complex system maintenance handbook. In: Kobbacy KAH, Murthy DNP (eds) Springer London, London, pp 263–286
33. Ramirez-Marquez JE, Coit DW (2007) Optimization of system reliability in the presence of common cause failures. Reliab Eng Syst Saf 92(10):1421–1434
34. Ye Z, Revie M, Walls L (2014) A load sharing system reliability model with managed component degradation. IEEE Trans Reliab 63(3):721–730
35. Hassett TF, Dietrich DL, Szidarovszky F (1995) Time-varying failure rates in the availability and reliability analysis of repairable systems. IEEE Trans Reliab 44(1):155–160

36. Bukowski JV, Goble WM (2001) Defining mean time-to-failure in a particular failure-state for multi-failure-state systems. IEEE Trans Reliab 50(2):221–228
37. Hougaard P (1989) Fitting a multivariate failure time distribution. IEEE Trans Reliab 38(4):444–448
38. Eryilmaz S (2011) Estimation in coherent reliability systems through copulas. Reliab Eng Syst Saf 96(5):564–568
39. Xi Z, Jing R, Wang P, Hu C (2014) A copula-based sampling method for data-driven prognostics. Reliab Eng Syst Saf 132:72–82
40. Shahraki AF, Yadav OP, Liao H (2017) A review on degradation modelling and its engineering applications. Int J Performability Eng 13(3):299
41. Alaswad S, Xiang Y (2017) A review on condition-based maintenance optimization models for stochastically deteriorating system. Reliab Eng Syst Saf 157:54–63
42. Wang H (2002) A survey of maintenance policies of deteriorating systems. Eur J Oper Res 139(3):469–489
43. Nicolai RP, Dekker R (2008) Optimal maintenance of multi-component systems: a review. Complex system maintenance handbook, pp 263–286
44. Keizer MCO, Flapper SDP, Teunter RH (2017) Condition-based maintenance policies for systems with multiple dependent components: a review. Eur J Oper Res 261(2):405–420
45. Bian L, Gebraeel N (2014) Stochastic framework for partially degradation systems with continuous component degradation-rate-interactions. Naval Research Logistics (NRL) 61(4):286–303
46. Hong H-P, Zhou W, Zhang S, Ye W (2014) Optimal condition-based maintenance decisions for systems with dependent stochastic degradation of components. Reliab Eng Syst Saf 121:276–288
47. Jafary B, Nagaraju V, Fiondella L (2017) Impact of correlated component failure on preventive maintenance policies. IEEE Trans Reliab 66(2):575–586
48. Bian L, Gebraeel N (2014) Stochastic modeling and real-time prognostics for multi-component systems with degradation rate interactions. IIE Trans 46(5):470–482
49. Rasmekomen N, Parlikad AK (2016) Condition-based maintenance of multi-component systems with degradation state-rate interactions. Reliab Eng Syst Saf 148:1–10
50. Li H, Deloux E, Dieulle L (2016) A condition-based maintenance policy for multi-component systems with Lévy copulas dependence. Reliab Eng Syst Saf 149:44–55
51. Xu D, Wei Q, Elsayed EA, Chen Y, Kang R (2017) Multivariate degradation modeling of smart electricity meter with multiple performance characteristics via vine copulas. Qual Reliab Eng Int 33(4):803–821
52. Dao CD, Zuo MJ (2015) Selective maintenance for multistate series systems with s-dependent components. IEEE Trans Reliab 65(2):525–539
53. Jafary B, Fiondella L (2016) A universal generating function-based multi-state system performance model subject to correlated failures. Reliab Eng Syst Saf 152:16–27
54. Feo TA, Bard JF (1989) Flight scheduling and maintenance base planning. Manage Sci 35(12):1415–1432
55. Cassady CR, Pohl EA, Murdock WP (2001) Selective maintenance modeling for industrial systems. J Qual Maintenance Eng. 7(2):104–117
56. Feng Q, Jiang L, Coit DW (2016) Reliability analysis and condition-based maintenance of systems with dependent degrading components based on thermodynamic physics-of-failure. Int J Adv Manuf Technol 86(1):913–923
57. Zhuang X, Yadav OP (2015) A new reliability assessment model for power electronic modules. In: 2015 IEEE international conference on industrial engineering and engineering management (IEEM). IEEE
58. Cafaro G, Corsi F, Vacca F (1986) Multistate Markov models and structural properties of the transition-rate matrix. IEEE Trans Reliab 35(2):192–200
59. Jianan X, Kai Y (1995) Dynamic reliability analysis of coherent multistate systems. IEEE Trans Reliab 44(4):683–688

60. Wenjian L, Hoang P (2005) Reliability modeling of multi-state degraded systems with multi-competing failures and random shocks. IEEE Trans Reliab 54(2):297–303
61. Wenjian L, Hoang P (2005) An inspection-maintenance model for systems with multiple competing processes. IEEE Trans Reliab 54(2):318–327
62. Kijima M, Nakagawa T (1991) A cumulative damage shock model with imperfect preventive maintenance. Naval Res Logistics (NRL) 38(2):145–156
63. Zhang YL, Yam RCM, Zuo MJ (2002) Optimal replacement policy for a multistate repairable system. J Oper Res Soc 53(3):336–341
64. Zhang YL, Yam RCM, Zuo MJ (2007) A bivariate optimal replacement policy for a multistate repairable system. Reliab Eng Syst Saf 92(4):535–542
65. Guo C, Wang W, Guo B, Si X (2013) A maintenance optimization model for mission-oriented systems based on Wiener degradation. Reliab Eng Syst Saf 111:183–194
66. Kijima M (1989) Some results for repairable systems with general repair. J Appl Probab 26(1):89–102
67. Yeh L (1988) A note on the optimal replacement problem. Adv Appl Probab 20(2):479–482
68. Liu Y, Huang HZ (2010) Optimal replacement policy for multi-state system under imperfect maintenance. IEEE Trans Reliab 59(3):483–495
69. Shahraki AF, Yadav OP, Vogiatzis C (2020) Selective maintenance optimization for multi-state systems considering stochastically dependent components and stochastic imperfect maintenance actions. Reliab Eng Syst Saf 196:106738
70. Wang GJ, Zhang YL (2005) A shock model with two-type failures and optimal replacement policy. Int J Syst Sci 36(4):209–214
71. Chien Y-H, Sheu S-H, Zhang ZG, Love E (2006) An extended optimal replacement model of systems subject to shocks. Eur J Oper Res 175(1):399–412
72. Nakagawa T (2007) Shock and damage models in reliability theory. Springer Science & Business Media
73. Yu L, Hong-Zhong H, Pham H (2008) Reliability evaluation of systems with degradation and random shocks. In: 2008 annual reliability and maintainability symposium (RAMS)
74. Peng H, Feng Q, Coit DW (2010) Reliability and maintenance modeling for systems subject to multiple dependent competing failure processes. IIE Trans 43(1):12–22
75. Song S, Coit DW, Feng Q, Peng H (2014) Reliability analysis for multi-component systems subject to multiple dependent competing failure processes. IEEE Trans Reliab 63(1):331–345
76. Song S, Coit DW, Feng Q (2014) Reliability for systems of degrading components with distinct component shock sets. Reliab Eng Syst Saf 132:115–124
77. Song S, Coit DW, Feng Q (2016) Reliability analysis of multiple-component series systems subject to hard and soft failures with dependent shock effects. IIE Trans 48(8):720–735
78. Yousefi N, Coit DW, Song S, Feng Q (2019) Optimization of on-condition thresholds for a system of degrading components with competing dependent failure processes. Reliab Eng Syst Saf 192:106547
79. Lu L, Wang B, Hong Y, Ye Z (2021) General path models for degradation data with multiple characteristics and covariates. Technometrics 63(3):354–369
80. Ye X, Hu Y, Zheng B, Chen C, Zhai G (2022) A new class of multi-stress acceleration models with interaction effects and its extension to accelerated degradation modelling. Reliab Eng Syst Saf 228:108815
81. McPherson JW (2018) Reliability physics and engineering: time-to-failure modeling. Springer, Berlin
82. Zhou S, Tang Y, Xu A (2021) A generalized Wiener process with dependent degradation rate and volatility and time-varying mean-to-variance ratio. Reliab Eng Syst Saf 216
83. Mercier S, Pham HH (2012) A preventive maintenance policy for a continuously monitored system with correlated wear indicators. Eur J Oper Res 222(2):263–272
84. Cherian KC (1941) A bi-variate correlated gamma-type distribution function. J Indian Math Soc 5:133–144
85. Liu B, Pandey MD, Wang X, Zhao X (2021) A finite-horizon condition-based maintenance policy for a two-unit system with dependent degradation processes. Eur J Oper Res 295(2):705–717

86. Liu T, Pan Z, Sun Q, Feng J, Tang Y (2017) Residual useful life estimation for products with two performance characteristics based on a bivariate Wiener process. Proc Inst Mech Eng Part O J Risk Reliab 231(1):69–80
87. Dong Q, Cui L, Si S (2020) Reliability and availability analysis of stochastic degradation systems based on bivariate Wiener processes. Appl Math Model 79:414–433
88. Nelsen RB, Krickeberg K, Fienberg S (1999) An introduction to copulas. Springer, Berlin
89. Pan Y, Ou SW, Zhang LM, Zhang WJ, Wu XG, Li H (2019) Modeling risks in dependent systems: a Copula-Bayesian approach. Reliab Eng Syst Saf 188:416–431
90. Yang Q, Zhang N, Hong Y (2013) Reliability analysis of repairable systems with dependent component failures under partially perfect repair. IEEE Trans Reliab 62(2):490–498
91. Liu XD, Pan F, Cai WL, Peng R (2020) Correlation and risk measurement modeling: a Markov-switching mixed Clayton copula approach. Reliab Eng Syst Saf 197
92. Hsu TM, Emura T, Fan TH (2016) Reliability inference for a copula-based series system life test under multiple type-I censoring. IEEE Trans Reliab 65(2):1069–1080
93. Li H, Zhu W, Dieulle L, Deloux E (2022) Condition-based maintenance strategies for stochastically dependent systems using Nested Lévy copulas. Reliab Eng Syst Saf 217:108038
94. Fang G, Pan R, Hong Y (2020) Copula-based reliability analysis of degrading systems with dependent failures. Reliab Eng Syst Saf 193:106618
95. Liu B, Zhao X, Liu G, Liu Y (2020) Life cycle cost analysis considering multiple dependent degradation processes and environmental influence. Reliab Eng Syst Saf 197:106784
96. Song K, Cui L (2022) A common random effect induced bivariate gamma degradation process with application to remaining useful life prediction. Reliab Eng Syst Saf 219:108200
97. Xu A, Shen L, Wang B, Tang Y (2018) On modeling bivariate wiener degradation process. IEEE Trans Reliab 67(3):897–906
98. Hong Y, Zhang M, Meeker WQ (2018) Big data and reliability applications: the complexity dimension. J Qual Technol 50(2):135–149
99. Yousefi N, Coit DW, Song S (2020) Reliability analysis of systems considering clusters of dependent degrading components. Reliab Eng Syst Saf 202:107005
100. Wu X, Huang T, Zhou K, Dai W (2022) Degradation modeling and reliability prediction of products with indicators influenced by clusters in a dynamic environment. Proc Inst Mech Eng Part O J Risk Reliab. https://doi.org/10.1177/1748006X221083417

A Study of Health State Transitions for Proactive Health Management

Lirong Cui, Weixin Jiang, Mengqian Wang, and He Yi

Abstract The proactive health management is a new medical mode which is becoming an important issue for national health management. Proactive health management is very similar to reliability, the health state of an individual or a group of people is the key issue. Thus, it is significant to focus on the related researches in both theory and applications. In the chapter, a new Markov process is developed for describing the evolution process of health states via considering the health state itself and the invention events, and based on the Markov model, the formulas for several related measure indexes in health states are derived. Meanwhile, some analyses on the sensitivity of parameters appeared in the model and some numerical examples to illustrate the related measure indexes are presented. This research may shed light on further studies in proactive health management.

Keywords Proactive health management · Health state · Transition · Markov process · Probability

1 Introduction

The health issue becomes much significant as our society developed rapidly. On the other hand, as science and technology development, the ideas and attitudes for people's health is changing. The traditional medicine faces some huge challenges, many patients cannot get their desired treatments. Thus, some new ideas, modes and medicine emerges, in which a so-called "proactive health management" has

L. Cui (✉) · W. Jiang · M. Wang
College of Quality and Standardization, Qingdao University, Qingdao 266071, China
e-mail: Lirongcui@qdu.edu.cn

L. Cui
China Sport Information Center, Beijing 100061, China

H. Yi
School of Economics and Management, Beijing University of Chemical Technology, Beijing, China

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_20

been proposed [1, 2]. Its main ideas are that it is a medical model to realize the body enhancement mechanism or chronic disappear via inputting proactively some controllable stimulations which promote the body's micro complexity increasing and increase the body's various adaptability. In proactive health management, the health states can be classified into several categories. For example, in [1], it proposed a body behavior mode: PAMS model (Pose, Action, Motion, Style) and proactive health model: SAEE model (Stimulation, Adaption, Evolution, Emergence). It is clear that the classification of health states to distinguish and measure the different health states, which are useful for the proactive health management, thus the related studies are significant in both theory and practice.

In the proactive health management, the stimulations can be treated as intervention events which may include physical excises, healthy diets, good behaviors and other healthy actions, which can also be thought of as shocks like in reliability field. Because of the diversity of each individual or a group of people, such as age, habitus, gender and other individual physical features, we treat them a system with some features. In fact, human body is a huge system which is more complex than our image. In this chapter, we shall mimic the way used in reliability to study the classification of proactive health states via using a finite state time homogenous Markov process. Meanwhile, the aggregated stochastic Markov processes are also used in the Chapter, which is a powerful tool to handle some problems.

The related literature is reviewed simply as follows. For proactive health management, there are few papers, which included Li's PhD dissertation [1, 2], both in Chinese, but it has not found some English literature on proactive health management. For the related Markov processes, there are many papers, for example, Zheng et al. [3] studied a single-unit Markov repairable system with repair time omission, Bao and Cui [4] considered a series repairable system with neglected or delayed failures, Cui et al. [5] studied a single unit system with state aggregation, Rubino and Sericola [6] considered the sojourn times in finite Markov processes. For aggregated stochastic Markov processes and their applications in reliability, there are a lot works, for example, Colquhoun and Hawkes [7], Cui et al. [8, 9], Hawkes et al. [10] and Yin and Cui [11]. For multi-state reliability systems, the related literature also has many, for example, Eryilmaz [12] and Wu et al. [13]. For shock models, there are much literature, for example, Cha and Finkelstin [14], Cui and Kang [15] and Yin and Cui [11]. All literature mentioned above will be valuable for our reference in this chapter work.

The contributions of this Chapter are: (i) To develop a new Markov process for describing the evolution process of health states via considering the health state itself and the invention events, (ii) to derive the formulas for serval related measure indexes in health states, (iii) to give some simple analyses on the sensitivity of parameters appeared in the model and, (iv) to present some numerical examples to illustrate the related measure indexes and some parameter effects on the results.

The reminder of the chapter is organized as follows. In Sect. 2, the assumptions to be needed in the development model are presented, then the reasons why these assumptions are raised are discussed simplify. The modelling and analyses are done in Sect. 3, in this section, a new Markov process is proposed to describe the evolution

of the health states, in which two processes are used, one is Markov process, another marked renewal process. After modelling, several measure indexes are considered, then their computation formulas are derived. In Sect. 4, a sensitivity of parameters is considered simply. A numerical example is presented to illustrate the results obtained in the Chapter in Sect. 5, in which include some sensitivity analysis. Finally, a conclusion is given in Sect. 6.

2 Assumptions

We have the following assumptions which will be used throughout the chapter.

(1) Suppose the evolution process of health states, denoted as $\{X(t), t \geq 0\}$, follow the time homogenous finite state Markov process with transition rate matrix $\mathbf{Q}_X$ and state space $S_X = \{0, 1, \ldots, m\}$ without any intervention acts.

(2) The intervention events, denoted as $\{N(t), t \geq 0\}$, arrive according to a renewal process with rate $\mu(\mu > 0)$ with marks $\{Y(t), t \geq 0\}$, the marks indicating the effect amounts of health states are either 1 or 2 with corresponding probabilities p and $1 - p$, respectively. That is to say, where $t_1 < t_2 < \cdots < t_n < \cdots$ are the renewal instants, $p \in [0, 1]$

$$Y(t_i) = \begin{cases} 1, & \text{with probability } p, \\ 2, & \text{with probability } 1 - p, \end{cases}$$

(3) Both stochastic processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ are independent with each other.

(4) The initial state of $\{X(t), t \geq 0\}$ is $\boldsymbol{\alpha}_0 = (\alpha_0, \alpha_1, \ldots, \alpha_m)$, and it is assumed that state i is better than state j if $i < j$, i.e., the states of $\{X(t), t \geq 0\}$ get worse as the labeling number of states increases.

(5) The transition rate matrix $\mathbf{Q}_X$ is given by where $\lambda_i > 0$, $\lambda_{ij} > 0$, $(i, j \in S_X)$, this implies that, if without any intervention, the health state can only be transferred from a better state to a worse state, and only the two nearest worse states can be reached directly after one transition.

$$\mathbf{Q}_X = \begin{pmatrix} -\lambda_0 & \lambda_{01} & \lambda_{02} & & \mathbf{0} \\ & -\lambda_1 & \lambda_{12} & \ddots & \\ & & \ddots & \ddots & \lambda_{(m-2)m} \\ & & & \ddots & \lambda_{(m-1)m} \\ \mathbf{0} & & & & -\lambda_m \end{pmatrix},$$

(6) The intervention act can improve the health state, and it is assumed that the health state process, denoted as $\{H(t), t \geq 0\}$.

The reasons why these assumptions are used are: (i) The evolution process is described, in general, with discrete states (multi-state), via a Markov process, this is

way the $\{X(t), t \geq 0\}$ is used to describe the evolution of health states without any intervention. And the transitions are limited at most two states, which may be coincident with real world situations without jump transitions or small step transitions. (ii) The invention events, described via a marked renewal process with a constant rate, and marks may be 1 and 2, which corresponds to the stimulations in health states cannot be jumped or small jumps. (iii) Other assumptions are very natural to fit real world situation in proactive health management and simple model consideration.

3 Modelling and Analyses

The evolution process of the health state transitions can be modeled by a finite state time homogeneous Markov process. Based on the assumptions, we can get that this stochastic process is $\{H(t), t \geq 0\}$ such that

$$H(t) \stackrel{d}{=} \begin{cases} X(t), \text{ with initial states } \boldsymbol{\alpha}_0, \text{ when } t < t_1, \\ X(t - t_1), \text{ with initial state } [X(t_1) - Y(t_1)] \vee 0, \text{ when } t_1 \leq t < t_2, \\ X(t - t_2), \text{ with initial state } [X(t_2 - t_1) - Y(t_2)] \vee 0, \text{ when } t_2 \leq t < t_3, \\ \cdots \\ X(t - t_n), \text{ with initial state } [X(t_n - t_{n-1}) - Y(t_n)] \vee 0, \text{ when } t_n \leq t < t_{n+1}, \\ \cdots, \end{cases}$$

where the instants $t_1 < t_2 < \cdots < t_n < \cdots$ are renewal points of $\{N(t), t \geq 0\}$, the symbol $\stackrel{d}{=}$ indicates the same distribution on both sides, the symbol $\vee$ is maximal operator, i.e., $a \vee b = \max\{a, b\}$. The $\{H(t), t \geq 0\}$ has a state space $S = \{0, 1, \ldots, m\}$. It is easy to know that the transition diagram of $\{H(t), t \geq 0\}$ is given by Fig. 1.

The transition rate matrix for $\{H(t), t \geq 0\}$ is

$$\mathbf{Q} = \mathbf{Q}_X + \mathbf{Q}_Y,$$

where

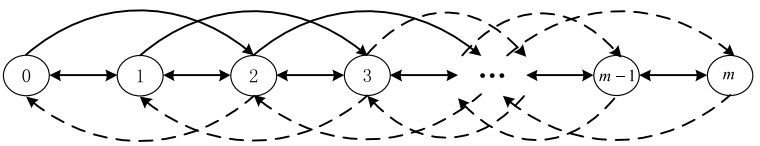


Fig. 1 The transition diagram of $\{H(t), t \geq 0\}$

$$\mathbf{Q}_Y = \begin{pmatrix} 0 & & & & & \mathbf{0} \\ \mu & -\mu & & & & \\ (1-p)\mu & p\mu & -\mu & & & \\ & (1-p)\mu & p\mu & -\mu & & \\ & & \ddots & \ddots & \ddots & \\ \mathbf{0} & & & (1-p)\mu & p\mu & -\mu \end{pmatrix}.$$

This is because we can treat the process $\{Y(t), t \geq 0\}$ as a time homogenous Markov process, and its role is to improve the health state of $\{X(t), t \geq 0\}$.

Then we have

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_{01} & \lambda_{02} & & & \mathbf{0} \\ \mu & -(\mu+\lambda_1) & \lambda_{12} & \lambda_{13} & & \\ (1-p)\mu & p\mu & -(\mu+\lambda_2) & \lambda_{23} & \ddots & \\ & (1-p)\mu & p\mu & -(\mu+\lambda_3) & \ddots & \lambda_{(m-2)m} \\ & & \ddots & \ddots & \ddots & \lambda_{(m-1)m} \\ \mathbf{0} & & & (1-p)\mu & p\mu & -(\mu+\lambda_m) \end{pmatrix}$$

The initial states of $\{H(t), t \geq 0\}$ is the same as that of $\{X(t), t \geq 0\}$, i.e., $\boldsymbol{\beta}_0 = \boldsymbol{\alpha}_0 = (\alpha_0, \alpha_1, \ldots, \alpha_m)$.

In the following, we shall present some analyses and computation formulas on the health state process $\{H(t), t \geq 0\}$.

(1) The instantaneous probability that the health state stays in a given set $\mathbf{S}_1 = \{0, 1, \ldots, n_1\}$ at time t.

Based on the theory of Markov process, we have

$$P\{H(t) \in \mathbf{S}_1\} = \boldsymbol{\beta}_0 \exp(\mathbf{Q}t)\tilde{\boldsymbol{u}}_1^T,$$

where the vector $\tilde{\boldsymbol{u}}_1 = (\underbrace{1, \ldots, 1}_{n_1}, 0, \ldots, 0),$, the superscript T is a transpose operator.

(2) The instantaneous probability that the health state stays in a given set $\mathbf{S}_1$ at time t and at most sojourns at set $\mathbf{S}_2 = \mathbf{S}/\mathbf{S}_1 = \{n_1 + 1, \ldots, m\}$ one time Based on the definition of the probability, we can know that the possible paths of health state by time t is shown in Fig. 2.

Thus, we have

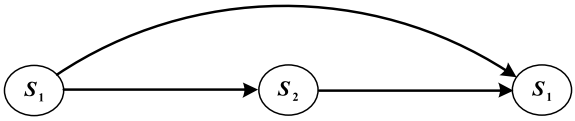


Fig. 2 The possible paths of $\{H(t), t \geq 0\}$

$$\varphi_1(t) := P\{H(t) \in \mathbf{S}_1,\, H(u) \text{ stays in } \mathbf{S}_2 \text{ for at most one time when } 0 < u < t\}$$
$$= \boldsymbol{\beta}_1 \exp(\mathbf{Q}_{11}t)\mathbf{u}_1^T + \boldsymbol{\beta}_1 \left[\iint\limits_{0\le u+v\le t} \mathbf{G}_{12}(u)\mathbf{G}_{21}(v) \exp(\mathbf{Q}_{11}(t-u-v))\mathrm{d}u\mathrm{d}v \right] \mathbf{u}_1^T,$$

where the vector $\boldsymbol{\beta}_1$ is the part elements with dimension $1 \times n_1$ such that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $\mathbf{u}_1 = (1, \ldots, 1)_{1\times n_1}$, the matrix $\mathbf{Q}_{11}$ with dimension $n_1 \times n_1$ is a submatrix of $\mathbf{Q}$, i.e.,

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix},$$

and $\mathbf{G}_{12}(t)$ is a matrix with dimension $n_1 \times n_2$ whose elements are defined as

$$G_{12}^{i,j}(t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} P\{H(t) \text{ stays in } \mathbf{S}_1 \text{ from time 0 to time } t, \text{ and leaves } \mathbf{S}_1$$
$$\text{forstate } j \text{ between } t \text{ and } t+\Delta t | H(0) = i\}, \quad i \in \mathbf{S}_1,\, j \in \mathbf{S}_2.$$

$\mathbf{G}_{21}(t)$ is a similar quantity resulting from $\mathbf{G}_{12}(t)$ by replacing 1 and 2 by 2 and 1, respectively. Both $\mathbf{G}_{12}(t)$ and $\mathbf{G}_{21}(t)$ are probability density functions for the sojourn times in subsets $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively, for details of these contents, see Colquhoun and Hawkes [7].

The convolution operation may be difficult for the related calculations directly, but we can use the Laplace transform to express and compute them concisely and simply. The Laplace transform of $\mathbf{G}_{12}(t)$(or $\mathbf{G}_{21}(t)$) is

$\mathbf{G}_{12}^*(s) = (s\mathbf{I} - \mathbf{Q}_{11})^{-1}\mathbf{Q}_{12}$, $(\mathbf{G}_{21}^*(s) = (s\mathbf{I} - \mathbf{Q}_{22})^{-1}\mathbf{Q}_{21})$,

where $\mathbf{I}$ is an identity matrix with proper dimension.

Furthermore, we have the Laplace transform of $\varphi_1(t)$ as

$$\varphi_1^*(s) = \boldsymbol{\beta}_1[(s\mathbf{I} - \mathbf{Q}_{11})^{-1} + \mathbf{G}_{12}^*(s)\mathbf{G}_{21}^*(s)(s\mathbf{I} - \mathbf{Q}_{11})^{-1}]\mathbf{u}_1^T$$
$$= \boldsymbol{\beta}_1[(s\mathbf{I} - \mathbf{Q}_{11})^{-1} + (s\mathbf{I} - \mathbf{Q}_{11})^{-1}\mathbf{Q}_{12}(s\mathbf{I} - \mathbf{Q}_{22})^{-1}\mathbf{Q}_{21}(s\mathbf{I} - \mathbf{Q}_{11})^{-1}]\mathbf{u}_1^T$$
$$= \boldsymbol{\beta}_1(s\mathbf{I} - \mathbf{Q}_{11})^{-1}[\mathbf{I} + \mathbf{Q}_{12}(s\mathbf{I} - \mathbf{Q}_{22})^{-1}\mathbf{Q}_{21}(s\mathbf{I} - \mathbf{Q}_{11})^{-1}]\mathbf{u}_1^T.$$

(3) The Instantaneous probability at state i ($i \in \mathbf{S}$)

Similar to the results in (1), we have that the instantaneous probability at state i is given by

$$\eta_i(t) := P\{H(t) = i\} = \boldsymbol{\beta}_0 \exp(\mathbf{Q}t)\tilde{\boldsymbol{u}}_{1,i}^T,$$

where $\tilde{\boldsymbol{u}}_{1,i} = (0, \ldots, 0, \overbrace{1}^{i\text{th}}, 0, \ldots, 0)$,

Similarly, the Laplace transform can be used for computation easily. Thus we have

$$\eta_i^*(s) = \boldsymbol{\beta}_0(s\mathbf{I} - \mathbf{Q})^{-1}\tilde{\boldsymbol{u}}_{1,i}^T,$$

(4) The steady-state probability

When looking at the health state of an individual or a group people during a long period of time, we can use the steady-state probability. i.e.,

$$\pi_i = \lim_{t\to\infty} P\{H(t) = i\},\ i \in S.$$

It is well-known that the steady-state probabilities satisfy the following linear set of equations, if let $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_m)^T$,

$$\begin{cases} \mathbf{Q}^T\boldsymbol{\pi} = \mathbf{0}, \\ \sum_{i=0}^{m} \pi_i = 1, \end{cases}$$

which is equivalent to the following linear set of equations

$$\tilde{\boldsymbol{Q}}_1\boldsymbol{\pi} = \mathbf{B}_1$$

where the vector $\mathbf{B}_1 = (1, 0, \ldots, 0)^T$, and

$$\tilde{\boldsymbol{Q}}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ \lambda_{01} & -(\mu+\lambda_1) & p\mu & (1-p)\mu & & \mathbf{0} \\ \lambda_{02} & \lambda_{12} & -(\mu+\lambda_2) & p\mu & \ddots & \\ & \lambda_{13} & \lambda_{23} & -(\mu+\lambda_3) & \ddots & (1-p)\mu \\ & & \ddots & \ddots & \ddots & p\mu \\ \mathbf{0} & & & \lambda_{(m-2)m} & \lambda_{(m-1)m} & -(\mu+\lambda_m) \end{pmatrix}.$$

Based on Cramer rule, we have

$$\pi_i = \frac{\det[\tilde{\boldsymbol{Q}}_{1,i}]}{\det[\tilde{\boldsymbol{Q}}_1]},\ i = 0, 1, \ldots, m,$$

where $\tilde{\boldsymbol{Q}}_{1,i}$ is a matrix obtained via replacing the ith column of $\tilde{\boldsymbol{Q}}_1$ by the vector $\mathbf{B}_1$.

(5) The interval probability that the health state stays in a given subset $\mathbf{S}_1$ during the interval $[t, t+a]$ $(a \geq 0)$

Let $A([t, t+a]) = P\{H(u) \in \mathbf{S}_1,$ for any $t \leq u \leq t+a\}$. Then based on the theory of aggregated Markov processes, we have

$$A^*([t, t+a])(s)$$

$$
\begin{aligned}
&= \boldsymbol{\alpha}_1 \sum_{r=0}^{\infty} \left[\mathbf{G}^*_{\mathbf{S}_1\mathbf{S}_2}(s)\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s)\right]^r \mathbf{P}^*_{\mathbf{S}_1\mathbf{S}_1}(s)\mathbf{P}_{\mathbf{S}_1\mathbf{S}_1}(a)\mathbf{u}_1^T + \boldsymbol{\alpha}_2\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s) \\
&\quad \sum_{r=0}^{\infty} \left[\mathbf{G}^*_{\mathbf{S}_1\mathbf{S}_2}(s)\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s)\right]^r \mathbf{P}^*_{\mathbf{S}_1\mathbf{S}_1}(s)\mathbf{P}_{\mathbf{S}_1\mathbf{S}_1}(a)\mathbf{u}_1^T \\
&= \boldsymbol{\alpha}_1 \left[\mathbf{I} - \mathbf{G}^*_{\mathbf{S}_1\mathbf{S}_2}(s)\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s)\right]^{-1} \mathbf{P}^*_{\mathbf{S}_1\mathbf{S}_1}(s)\mathbf{P}_{\mathbf{S}_1\mathbf{S}_1}(a)\mathbf{u}_1^T + \boldsymbol{\alpha}_2\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s) \\
&\quad \left[\mathbf{I} - \mathbf{G}^*_{\mathbf{S}_1\mathbf{S}_2}(s)\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s)\right]^{-1} \mathbf{P}^*_{\mathbf{S}_1\mathbf{S}_1}(s)\mathbf{P}_{\mathbf{S}_1\mathbf{S}_1}(a)\mathbf{u}_1^T \\
&= \left[\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s)\right]\left[\mathbf{I} - \mathbf{G}^*_{\mathbf{S}_1\mathbf{S}_2}(s)\mathbf{G}^*_{\mathbf{S}_2\mathbf{S}_1}(s)\right]^{-1} \mathbf{P}^*_{\mathbf{S}_1\mathbf{S}_1}(s)\mathbf{P}_{\mathbf{S}_1\mathbf{S}_1}(a)\mathbf{u}_1^T \\
&= \left[\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2(s\mathbf{I} - \mathbf{Q}_{22})^{-1}\mathbf{Q}_{21}\right]\left[\mathbf{I} - (s\mathbf{I} - \mathbf{Q}_{11})^{-1}\mathbf{Q}_{12}(s\mathbf{I} - \mathbf{Q}_{22})^{-1}\mathbf{Q}_{21}\right]^{-1} \\
&\quad (s\mathbf{I} - \mathbf{Q}_{11})^{-1} \exp(a\mathbf{Q}_{11})\mathbf{u}_1^T,
\end{aligned}
$$

where $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$.

The other way for the interval probability is

$$A([t, t+a]) = \boldsymbol{\beta}_1 \exp(\mathbf{Q}t)(\mathbf{I}, \mathbf{0}) \exp(a\mathbf{Q}_{11})\mathbf{u}_1^T.$$

(6) The limit interval probability that the health state stays in a given subset $\mathbf{S}_1$ during the interval $[t, t+a]$ $(a \geq 0)$

Let $A([a]) = \lim_{t\to\infty} A([t, t+a]) = P\{H(u) \in \mathbf{S}_1, \text{ for any } \infty \leq u \leq \infty + a\}$.

Thus we have

$$A([a]) = (\pi_0, \ldots, \pi_{|\mathbf{S}_1|}) \exp(a\mathbf{Q}_{11})\mathbf{u}_1^T.$$

(7) What is the condition for $p \in [0, 1]$ such that

$$\sum_{i=0}^{n_1} \pi_i > \rho_1 \text{ and } \sum_{i=n_1+1}^{m} \pi_i < \rho_2,$$

where the parameters $\rho_1, \rho_2 \in [0, 1]$.

It is clear that p meets the following equality

$$p* = \max_{p\in[0,1]} \{p : \quad \sum_{i=0}^{n_1} \pi_i > \rho_1 \text{ and } \sum_{i=n_1+1}^{m} \pi_i < \rho_2\},$$

This is because $\sum_{i=0}^{n_1} \pi_i$ decrease in p and $\sum_{i=n_1+1}^{m} \pi_i$ decreases too in p, although the rigorous proof needs much effort.

(8) Special cases

There are two special cases needed to be considered. (i) $p = 0$ and (ii) $p = 1$. For case of $p = 0$, the marked renewal process $\{Y(t), t \geq 0\}$ is with rate μ and mark 2

at each renewal point, which indicates the strongest invention events happened. For case of $p = 1$, the marked renewal process $\{Y(t), t \geq 0\}$ is with rate μ and mark 1 at each renewal point, which indicates the weakest invention events happened.

(9) Mean and variance of $H(t)$

The mean of $\{H(t), t \geq 0\}$ is given by

$$M_1(t) := \mathcal{E}[H(t)] = \sum_{i=0}^{m} i\, P\{H(t) = i\} = \sum_{i=0}^{m} i\, \eta_i(t)$$

and its Laplace transform is

$$M_1^*(s) = \sum_{i=0}^{m} i\, \eta_i^*(s) = \sum_{i=0}^{m} i \boldsymbol{\beta}_0 (s\mathbf{I} - \mathbf{Q})^{-1} \tilde{\boldsymbol{u}}_{1,i}^T.$$

The second moment of $\{H(t), t \geq 0\}$ is given by

$$M_2(t) := \mathcal{E}[H^2(t)] = \sum_{i=0}^{m} i^2 P\{H(t) = i\} = \sum_{i=0}^{m} i^2 \eta_i(t)$$
$$= \sum_{i=0}^{m} i^2 \boldsymbol{\beta}_0 (s\boldsymbol{I} - \boldsymbol{Q})^{-1} \tilde{\boldsymbol{u}}_{1,i}^T.$$

Thus the variance of $\{H(t), t \geq 0\}$ is given by

$$Var[H(t)] = M_2(t) - [M_1(t)]^2.$$

4 Sensitivity Analyses on Parameters

The model (Markov process) involves the following non-negative parameters listed in Table 1. There are total $2m + 3$ parameters in the model.

In the sequel, we shall simply discuss the sensitivity of parameters in two cases: (i) Parameters in the input intensity $\{N(t), t \geq 0\}$ and, (ii) parameters in the evolution process of health states.

Table 1 Parameters in the model

$Y(t)$	$X(t)$
$p,\ \mu$	$\lambda_{i(i+1)},\ \lambda_{i(i+2)},\ i = 0, 1, \ldots, m-2.$ $\lambda_{(m-1)m}$

Case 1. Parameters in the input intensity $\{N(t), t \geq 0\}$.

There are two parameters in $\{Y(t), t \geq 0\}$: p and μ. It is intuitively to know that the health states will be in better states when p decreases or μ increases under other parameters being fixed, which have been shown in the numerical example. The average input intensity is

$$p\mu + 2(1-p)\mu = 2\mu - p\mu = (2-p)\mu.$$

Similarly, if the average input intensity increases, then the health states will stay in better states. The special cases: $p = 0$ and $p = 1$, have been considered in Sect. 3 already.

Case 2. Parameters in the evolution process of health states.

There are $2m + 1$ parameters in $\{X(t), t \geq 0\}$. The detailed discussion for the sensitivity of parameters will be difficult and complicated. But if some $\lambda_{i(i+1)}$ decrease, then the health states will stay in better states under other parameters being fixed. The situations become complicated when some $\lambda_{i(i+1)}$ decrease and some $\lambda_{i(i+1)}$ increase, in which we need to give the results based on the numerical calculations.

5 Numerical Example

In this section, a numerical example is given to show the related results presented Sect. 3. It is assumed that $m = 5$, $p = 0.8$, $\mu = 5$, and the transition rate matrix of $\{X(t), t \geq 0\}$ is

$$\mathbf{Q}_X = \begin{pmatrix} -3 & 1 & 2 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 & 0 & 0 \\ 0 & 0 & -4 & 2 & 2 & 0 \\ 0 & 0 & 0 & -2 & 2 & 0 \\ 0 & 0 & 0 & 0 & -5 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

the initial state is $\boldsymbol{\beta}_0 = \boldsymbol{\alpha}_0 = (0.2, 0.1, 0.2, 0.4, 0.1, 0)$, $\mathbf{S}_1 = \{0, 1, 2\}$.

Using Maple software and the results presented in Sect. 2, we have

(1) The instantaneous probability in $\mathbf{S}_1$ at time t

$$\mathbf{Q}_Y = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \mu & -\mu & 0 & 0 & 0 & 0 \\ (1-p)\mu & p\mu & -\mu & 0 & 0 & 0 \\ 0 & (1-p)\mu & p\mu & -\mu & 0 & 0 \\ 0 & 0 & (1-p)\mu & p\mu & -\mu & 0 \\ 0 & 0 & 0 & (1-p)\mu & p\mu & -\mu \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & -5 & 0 & 0 & 0 & 0 \\ 1 & 4 & -5 & 0 & 0 & 0 \\ 0 & 1 & 4 & -5 & 0 & 0 \\ 0 & 0 & 1 & 4 & -5 & 0 \\ 0 & 0 & 0 & 1 & 4 & -5 \end{pmatrix}.$$

$$\mathbf{Q} = \mathbf{Q}_X + \mathbf{Q}_Y = \begin{pmatrix} -3 & 1 & 2 & 0 & 0 & 0 \\ 5 & -7 & 1 & 1 & 0 & 0 \\ 1 & 4 & -9 & 2 & 2 & 0 \\ 0 & 1 & 4 & -7 & 2 & 0 \\ 0 & 0 & 1 & 4 & -10 & 5 \\ 0 & 0 & 0 & 1 & 4 & -5 \end{pmatrix}.$$

The curve of $P\{H(t) \in \mathbf{S}_1\}$ is shown in Fig. 3.

(2) The instantaneous probability in $\mathbf{S}_1$ at time t with at most one sojourn in $\mathbf{S}_2$

The curve of $\varphi_1(t)$ is shown in Fig. 4.

(3) The Instantaneous probabilities at state $i (i \in \mathbf{S})$

The curves of $\eta_i(t)(i = 0, 1, \ldots, 5)$ are shown in Fig. 5.

(4) The steady-state probability

The steady-state probability of $\{H(t), t \geq 0\}$ is given by

$$\boldsymbol{\pi} = (0.3237, 0.1612, 0.1652, 0.1439, 0.1030, 0.1030).$$

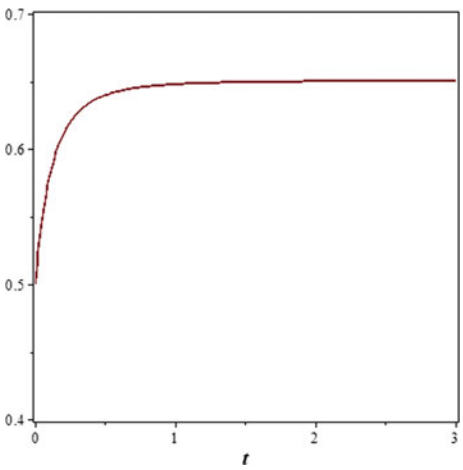


Fig. 3 The instantaneous probability in set $\mathbf{S}_1$ at time t

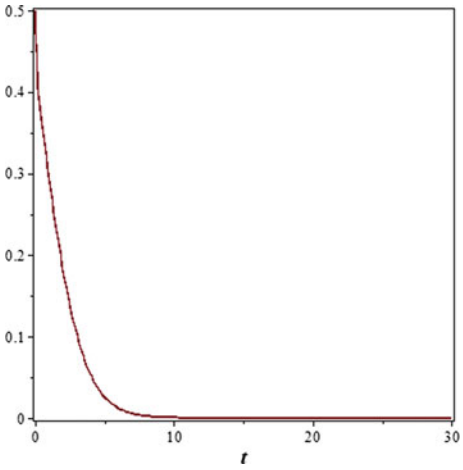


Fig. 4 The instantaneous probability in $\mathbf{S}_1$ at time t with at most one sojourn in $\mathbf{S}_2$

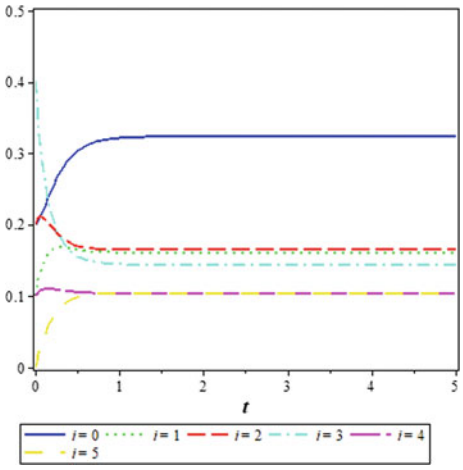


Fig. 5 The instantaneous probabilities at state $i (i \in \mathbf{S})$

(5) The interval probability in $\mathbf{S}_1$ during $[t, t + a]$

The curves of $A([t, t + a])$ are given in Fig. 6.

(6) The limit interval probability in $\mathbf{S}_1$ during $[t, t + a]$
(7) The impact of p on the limit probability in $\mathbf{S}_1$

The curve of the impact of p on the limit probability in $\mathbf{S}_1$ is given in Fig. 8.

(8) Mean and variance of $H(t)$

The curves of mean and variance are presented in Fig. 9, respectively.

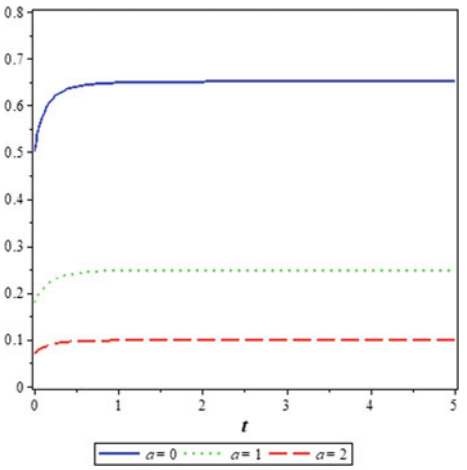


Fig. 6 The interval probabilities in $\mathbf{S}_1$ during $[t, t+a]$ for $a = 1, 2, 3$, respectively

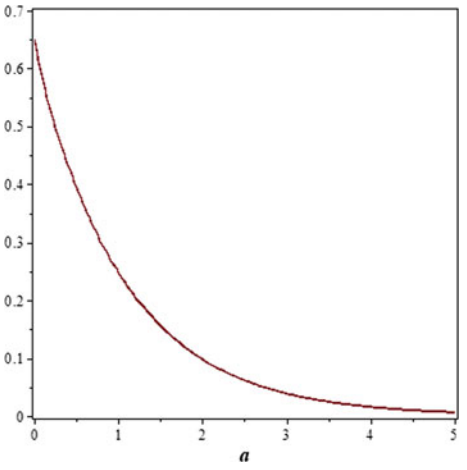


Fig. 7 The limit interval probability in $\mathbf{S}_1$ during $[t, t+a]$ for $a \in [0, 5]$

6 Conclusion

The evolution process of proactive health states via a new Markov process combining a Markov process and a marked renewal process, in which the quantitative transitions are described. Based on this new model, some measure indexes are proposed, and their calculation formulas are derived, some of them expressed via the Laplace transforms and some via matrix forms by using knowledge of Markov processes and aggregated Markov processes. A simple discussion on sensitivity of parameters appeared in the model are given, which may be useful in applications. Some numerical examples are presented to illustrate the results obtained in the chapter. The future related research may be done on semi-Markov process and other ways on intervention events such as dependent cases and complicated shock-like cases.

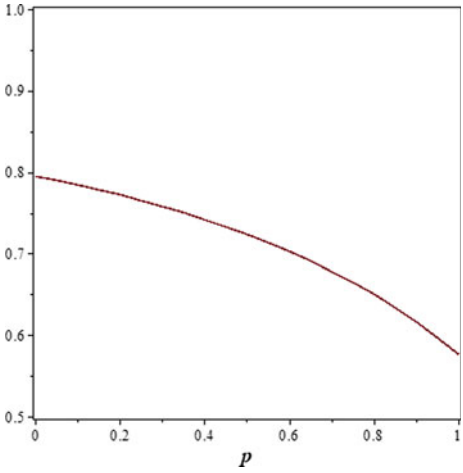


Fig. 8 The limit probability in $\mathbf{S}_1$ for $p \in [0, 1]$

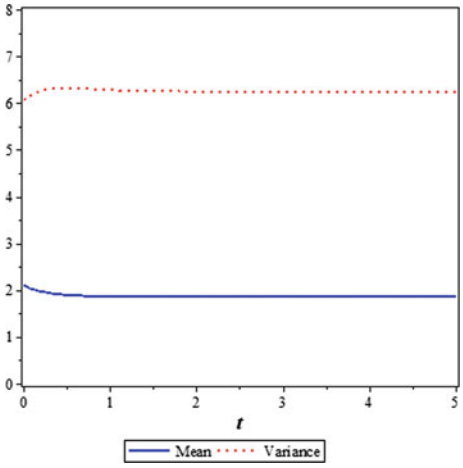


Fig. 9 Mean and variance of $H(t)$

Acknowledgements This work was supported by National Key R&D Program of China (2020YFC2007201) and National Natural Science Foundation of China under grants 72271134 and 71871021.

References

1. Li XC (2020) Research on the mechanism and measurement of proactive health behavior intervention. PhD dissertation. Beijing Institute of Technology (in Chinese)
2. Li XC, Yu MS (2020) Proactive health: concepts & modes. China Sport Sci 40:83–87 (in Chinese)

3. Zheng ZH, Cui LR, Hawkes AG (2006) A study on a single-unit Markov repairable system with repair time omission. IEEE Trans Reliab 55:182–188
4. Bao XZ, Cui LR (2010) An analysis of availability for series Markov repairable system with neglected or delayed failures. IEEE Trans Reliab 59:734–743
5. Cui LR, Du SJ, Hawkes AG (2012) A study on a single-unit repairable system with state aggregations. IIE Trans 44:1022–1032
6. Rubino G, Sericola B (1989) Sojourn times in finite Markov processes. J Appl Probab 26:744–756
7. Colquhoun D, Hawkes AG (1982) On the stochastic properties of bursts of single ion channel openings and of clusters of bursts. Philos Trans R Soc Lond Ser B: Biol Sci 300:1–59
8. Cui LR, Du SJ, Liu BL (2013) Multi-point and multi-interval availabilities. IEEE Trans Reliab 62:811–820
9. Cui LR, Du SJ, Zhang AF (2014) Reliability measures for two-part partition of states for aggregated Markov repairable systems. Ann Oper Res 212:93–114
10. Hawkes AG, Cui LR, Du SJ (2014) Occupancy times for Markov and semi-Markov models in systems. In: Frenkel IB, Karagrigoriou A, Lisnianski A, Kleyner A (eds) Reliability applied reliability engineering and risk analysis: probabilistic models and statistical inference. Wiley, Chichester, pp 218–230
11. Yin J, Cui LR (2022) Reliability analysis for shock systems based on damage evolutions via Markov processes. Naval Research Logistics 70(3):246–260.
12. Eryilmaz S (2013) On mean residual life of discrete time multi-state systems. Qual Technol Quant Manage 10:241–250
13. Wu B, Cui LR, Fang C (2020) Generalized phase-type distributions based on multi-state systems. IISE Trans 52:104–119
14. Cha JH, Finkelstein M (2018) Point processes for reliability analysis: shocks and repairable systems. Springer, New York
15. Cui LR, Kang FM (2022) Shocks models with damage effect evolutions following Markov processes. J Oper Res Soc 74(1):430–444.

Kalman Filter-Based Systems Approach for Prognostics and Health Management of Electric Motors

Hyung Jun Park, Dongwoo Lee, Seokgoo Kim, Nam Ho Kim, and Joo-Ho Choi

Abstract A Kalman filter-based framework is proposed for the prognostics and health management of DC electric motors by treating them as a system. The control signals of the motor are used to estimate the current health and predict the remaining useful life (RUL) of the motor and its components, such as bearings and permanent magnets. The framework consists of an online health diagnosis to estimate the health status of the motor and each component, and an offline failure prognosis to predict the RULs. The approach is demonstrated with the aid of two real examples: the reaction wheel motor for advanced attitude control of satellites and the driving motors in a quadcopter to lift and control flight operations. In each example, the motors were subjected to accelerated degradation tests, motor control data were collected for each cycle, and RULs were predicted against failure thresholds critical to motor performance. The results showed that the framework can be used to effectively predict the RUL of a degraded motor, thereby enabling failure prevention and proactive maintenance scheduling.

H. J. Park
Department of Aerospace and Mechanical Engineering, Korea Aerospace University, Gyeong gi-do 10540, Republic of Korea

D. Lee
D&SHINE Research Institute, Gyeong gi-do 10540, Republic of Korea

S. Kim · N. H. Kim
Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611, USA

J.-H. Choi (✉)
School of Aerospace and Mechanical Engineering, Korea Aerospace University, Gyeong gi-do 10540, Republic of Korea
e-mail: jhchoi@kau.ac.kr

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_21

1 Introduction

The accurate prediction of impending failure or remaining useful life (RUL) of mission-critical systems improves safe system operations and offers economic benefits to the industry. Extensive research on the prognostics and health management (PHM) of various assets with different perspectives has already been conducted by academic researchers and industrial engineers [1, 2]. Among these, the PHM of electric motors has been widely researched because of its ability to provide the drive and control of various equipment and processes in the industry. To conduct PHM, motor current or vibration signals are typically employed to assess and predict the health of a motor [3]. Traditionally, the failure modes of a motor have been identified by time and frequency analyses of the signals, such as by estimating the harmonic components of the fault frequencies and locating them in the spectrum [4–7]. Currently, artificial intelligence (AI)-based approaches, which are known to be powerful and possess improved performance over the conventional approaches, are widely adopted. These approaches consist of classification tools and algorithms, such as neural networks [8], fuzzy logic [9], support vector machines [10], and deep learning [11]. A general review of AI-based approaches for electric motors is presented in Ref. [12].

Most engineering systems consist of multiple components designed to perform a specific function. Because these components interact with each other in a complex manner, their degradation can affect the overall system performance in a non-trivial way. Recently, an appropriate framework that estimates the health of the components and predicts their RUL was proposed by Kim et al. [13, 14]. They also demonstrated the framework using a DC motor, considering it as a system because it has mechanical parts, that is, the bearing, shaft, and electrical components such as the stator winding and magnets. Because the performance degradation of the motor originates from these components, it is desirable to employ a systems approach to implement the PHM for motors. The framework consists of online diagnostics to monitor the status of each component and the overall health of the motor, as well as an offline failure prognosis to predict the RULs against the failure threshold conditional on motor performance.

Prognostic methods applied in PHM can be categorized into model-based and data-driven approaches. The model-based approach assumes that a physical model describing the behavior of the system is available and combines the model with measured data to identify the model parameters. Conversely, the data-driven approach uses data from past failures to establish a prediction model based on machine learning. The model-based approach is advantageous for DC motors because it is a dynamic model that enables estimation of the model parameters using the measured signals, which are indicative of the component's health. Hence, several researchers have applied model-based approaches to the PHM of DC motors. In this approach, the Kalman filter (KF) is the most widely used technique for estimating the model parameters in a recursive manner while incorporating the model and measurement uncertainties. Rahimi et al. [15] applied an unscented KF with a high-fidelity model to estimate health parameters for each motor fault scenario. El Sayed et al. [16] performed

parameter estimation using an extended KF (EKF) and an unscented KF (UKF) to diagnose stator faults and their severity. Others, such as Singleton et al. [17], used a KF to estimate motor speed and conducted fault detection based on a comparison with the actual speed. However, a common drawback of the aforementioned studies is that no one has explored the systems approach, wherein the RUL prediction accounts for component degradation and their influence on the overall motor performance.

This section addresses the application of the systems approach, in which the KF is used with motor-current signals to estimate the present health status of individual components and performance degradation. The advantages of the systems approach are two-fold. First, it does not require large volume of data until failure, which is critical to the training process in the data-driven approach. Second, additional sensors, such as accelerometers, are not required because the motor load current signal, which is acquired during operation, is used for the PHM. In Sect. 2, the overall prognostics framework for the systems approach is reviewed briefly. In Sect. 3, simulations are conducted using the motor dynamic equations in which the two most critical components, that is, the permanent magnet and bearing, are artificially degraded over cycles, and virtual measurements are performed accordingly to illustrate the process of RUL prediction. Next, two case studies of real motors are presented to demonstrate their implementation. The first is presented in Sect. 4, which considers the degradation of the reaction wheel (RW) motor, which stabilizes the attitude control of satellites against external environmental factors. The second is presented in Sect. 5 for the motors driving the quadcopter to lift and control flight operations. In both examples, the motors were subjected to accelerated degradation and the motor-current signals were collected with regular time interval during operation. A systems approach is employed wherein the performance of the motor is properly defined. The health of each motor component is assessed based on this, and the RUL, after which the motor is no longer able to perform its normal function, is predicted. In Sect. 6, a detailed discussion and conclusions are presented.

2 Systems Approach for PHM

The overall framework of the systems approach for fault diagnosis and failure prognosis based on a physical model-based approach is described in this section. Figure 1 describes the framework which consists of two phases: construction of the system dynamics model on the left and the PHM implementation on the right.

The system model is developed using simulation tools or algorithms such as Simulink, lumped parameters, or ordinary differential equations. A group of parameters is identified during development; the inputs are the operation parameter u and the health parameter h of the critical components that affect system degradation. The outputs are the state variable x, the system performance variable S, and the measurement variable z used to estimate the present health status of the system and its components. Three models are established: the state model—typically in a recursive form in the time domain—the measurement model relating the state with

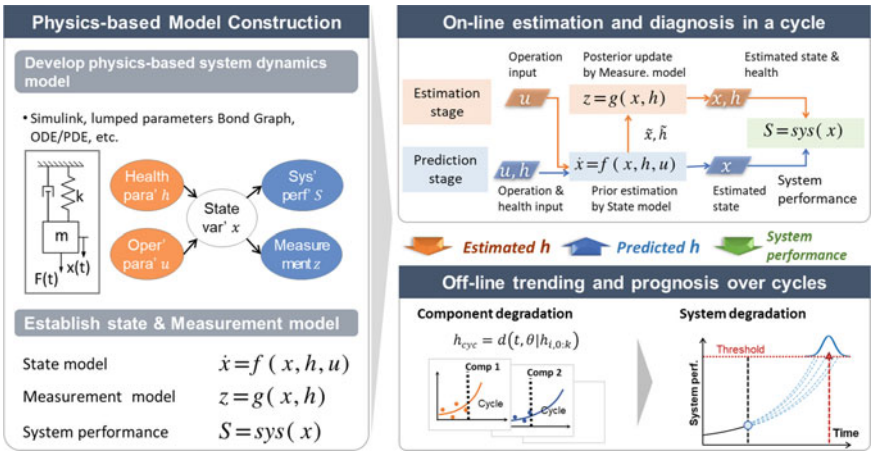


Fig. 1 Framework for physics-based approach [13]

the measurement, and the performance model of the system. The three models are defined as follows:

$$\dot{x} = f(x, h, u), \tag{1}$$

$$z = g(x, h), \tag{2}$$

$$S = sys(x). \tag{3}$$

The PHM implementation phase consists of two parts: online estimation and diagnosis in the upper-right figure, as well as offline training and prognosis in the lower-right figure. In the online estimation, state x in a single cycle is estimated for the input operation parameter u, from which the system performance S is evaluated using the performance model. Depending on the availability of health parameter h, the online estimation phase comprises two stages. In the estimation stage, h and x are estimated as unknowns using the measured data z. The unknown parameters are estimated by the state model and updated by the measurement model. In the prediction stage, h is known as a priori and there are no measurements; in this case, state x is predicted with the given input h. In both stages, the system performance is computed using the obtained state x. Using the measurement z obtained in every cycle, the health parameters up to the current cycle k, i.e., $h_{0:k}$, can be estimated. They are then transferred to offline training and prognosis. To describe the degradation trend more efficiently, the health parameters are typically represented by either a physical model or an empirical model:

$$h_{cyc} = d(t, \theta|h_{0:k}), \tag{4}$$

where h_{cyc} denotes the health parameters as a function of cycles, d is the mathematical model, θ is the model parameter, and t is the cycle. Once the model is fitted to $h_{0:k}$, it can be extrapolated to predict h in the future. The predicted h is transferred to the prediction stage from which state x and the system performance S in the future are predicted. The predicted S is transferred to the offline training phase to obtain the future evolution of the system performance, as matched with that of the health parameter h in the future. Upon evaluating the RUL of the system performance against the failure threshold, one can identify the most critical health parameter (i.e., specific component) that leads to the earliest system failure and its remaining cycles, which provides valuable information in maintenance management.

The overall procedure is illustrated in Fig. 2. For the implementation phase, the online estimation and diagnosis part of Fig. 1 can be best accommodated by a Bayesian approach, such as the Extended Kalman or Particle Filter algorithms. Thereafter, the health parameter h and state x are estimated in the form of a distribution, such as the mean and covariance, or the samples which reflect the uncertainty in the process. The offline training and prognosis parts can also be similarly performed; however, a simpler linear/nonlinear regression can also be employed to this end. The degradation model parameter θ in $d(\cdot)$ is estimated based on the accumulated values $h_{0:k}$ until the current cycle. The degradation in the future is then predicted by θ with uncertainty, which is usually expressed by the confidence bounds in the result.

Construction phase (left figure)	
• Establish state, measurement, and performance model, as given in (1)-(3)	
Implementation phase (right figure)	
• Estimate health parameter h and state x using the measurement z.	Estimation stage in upper figure
• Transfer estimated data $h_{0:k}$ up to the current cycle to offline trending.	
• Fit the data $h_{0:k}$ to degradation model (4) and extrapolate h to predict the future degradation.	Left part of lower figure
• Transfer predicted health parameters h in the future to online estimation.	
• Predict state x and system performance S using h in the future.	Prediction stage in upper figure
• Transfer predicted S in the future to offline trending.	
• Determine the remaining life of S against failure threshold.	Right part of lower figure

Fig. 2 Overall procedure for physics-based approach as illustrated

3 Simulation Study

3.1 Problem Statement

In this section, the simulation data are used to illustrate the system approach procedure for a DC motor. Most of the contents here are extracted from the authors' previous study [13]. Virtual measurement data are generated by adding random noise to the simulation results of the motor dynamics that converts electrical energy into mechanical power. When an electric current passes through a coil in a magnetic field, the magnetic force produces torque that drives the DC motor. As shown in Fig. 3, a DC motor consists of electrical and mechanical parts that are coupled together; failure in one part affects the others. For example, because magnetic flux enables conversion of electrical energy into mechanical force, its defects can degrade mechanical performance. The mechanical and electrical parts of the DC motor dynamics are expressed as:

$$J\frac{d\omega}{dt} + b\omega = k_T i - T_L = T_o, \tag{5}$$

$$L\frac{di}{dt} + Ri = V - k_T\omega, \tag{6}$$

where ω, i, and T_o represent the angular velocity, current, and output torque, respectively. The two equations share the common parameter k_T, which represents the electromechanical coupling coefficient, and Table 1 lists the parameters and their values used for the simulation [18]. The duration of a single cycle is 3 s, and a voltage of 10 V is applied for the first 1.5 s and subsequently turned off. Figure 4 shows the time histories of ω, i, and T_o as outputs from the dynamic equations in (5) and (6), respectively. The angular velocity ω rapidly reached the desired value on application of power and decreased to zero when turned off. Similarly, load current i and output torque T_o rapidly increased to the peak value at the start, followed by a gradual decrease when the voltage was turned off.

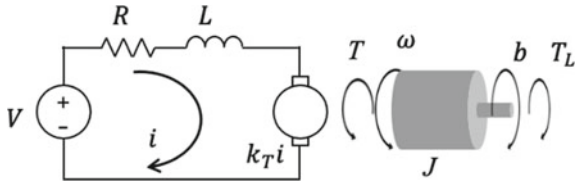


Fig. 3 DC motor system

Table 1 Parameter setting for simulation

Symbol	Description	Value
R	Armature resistance	11.2
L	Armature inductance	0.1215 H
J	Moment of inertia	0.022145 kg m^2
b	Viscous friction coefficient	0.002953 N m s/rad
k_T	Electromechanical coupling coefficient	1.28 Nm/A
V	Input voltage	10 V
T_L	Load torque	0.05 Nm

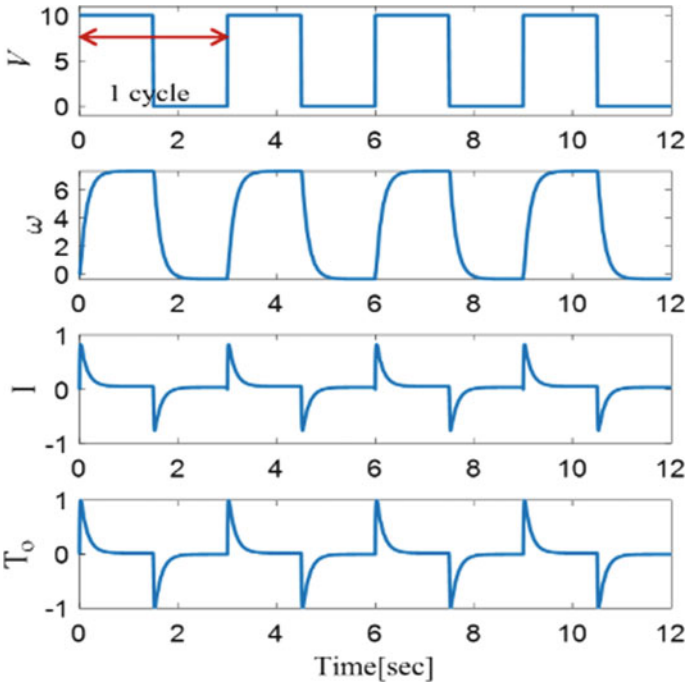


Fig. 4 Simulation result of DC motor system

3.2 Simulation of Component Degradation

Challenges with system-level prognosis arise because multiple components degrade over time, which affect system performance in a complex manner. To simulate this situation, two components with associated failure modes were selected from [19]. The first is degradation of the permanent magnet owing to prolonged overheating, also called flux weakening, that results in reduction of rotor magnetic-field strength. This can be described by decreasing the electromechanical coupling coefficient,

k_T. The second is bearing lubrication failure, which can be modeled by a change in the load torque T_L applied to the motor. The system performance is given by the motor output torque T_o, which may decrease as the two components degrade. The degradation behaviors of the magnet and bearing are assumed to be linear and exponential functions of the cycles, respectively.

$$k_T(t) = \alpha_1 + \alpha_2 t, \tag{7}$$

$$T_L(t) = \beta_1 e^{\beta_2 t}, \tag{8}$$

where α_1 and β_1 represent the initial degraded status, and α_2 and β_2 describe the cycle-dependent behavior.

Three cases are considered, as shown in Fig. 5: (1) degradation of the magnet with $\alpha_2 = -5.7974 \times 10^{-4}$, (2) degradation of the bearing with $\beta_2 = 3.3 \times 10^{-3}$, and (3) simultaneous degradation with the $\alpha_2 = -5.7974 \times 10^{-4}$ and $\beta_2 = 3.9 \times 10^{-3}$. The output torque for each case was obtained by solving the system equations with the degraded values of k_T and T_L, as shown in Fig. 5a–e; results for the output torque are presented in Fig. 5b, d, and f. It can be observed that the maximum value of the output torque gradually decreases with degradation, thereby indicating degradation of the system performance. Furthermore, it is evident from Fig. 5b, d, and f that when two or more components degrade simultaneously, the system performance degrades at a faster rate.

In this example, the system performance S is defined by a scalar value, namely the maximum output torque:

$$S = \max T_0 = \max(k_T i - T_L). \tag{9}$$

A system is regarded as a failure when the system performance decreases below 70% of its initial value. Therefore, it is necessary to predict the number of cycles that remain before failure, as well as the responsible component to be replaced.

3.3 Application of Physics-Based Approach

As mentioned in the previous section, the first step in the physics-based approach is construction of the system dynamics model, as shown in (1), where the state variable $\boldsymbol{x}$ (denoted by a vector hereafter) consists of velocity ω and current i, that is, $\boldsymbol{x} = [\omega, i]^T$ while the input operation parameter u is voltage V. The health parameter $\boldsymbol{h}$ contains k_T and T_L or $\boldsymbol{h} = [k_T, T_L]^T$. The state model can be constructed using (5) and (6) as follows:

$$\begin{bmatrix} \dot{\omega} \\ i \end{bmatrix} = \begin{bmatrix} -\frac{b}{J} & \frac{k_T}{J} \\ \frac{k_T}{L} & -\frac{R}{L} \end{bmatrix} \begin{bmatrix} \omega \\ i \end{bmatrix} + \begin{bmatrix} \frac{T_L}{J} \\ \frac{V}{L} \end{bmatrix}. \tag{10}$$

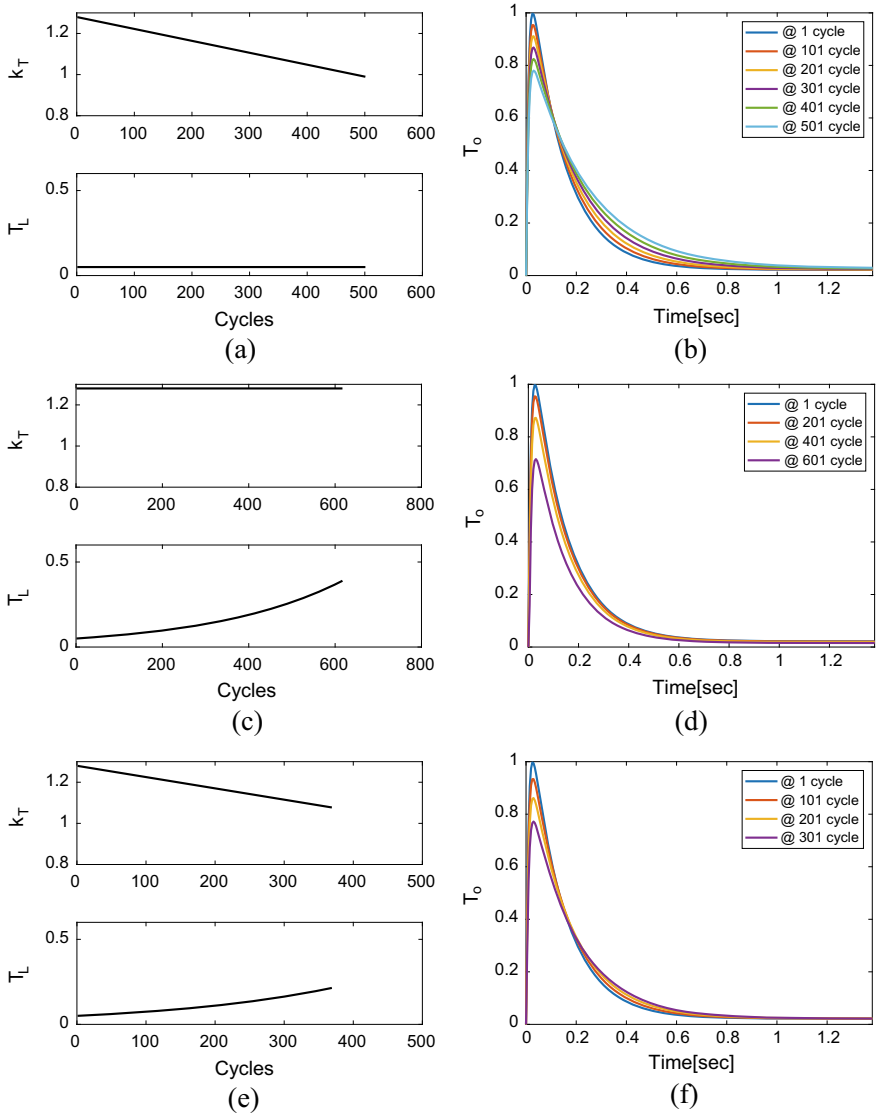


Fig. 5 Output torque due to components degradation, **a** health parameters and **b** output torque of case 1, **c** health parameters and **d** output torque of Case 2, and **e** health parameters and **f** output torque of Case 3

Note that the equation describes the transient response of the state variable $\boldsymbol{x}$. Because the state variable can be acquired from the control unit during operation, the measurement model is expressed as:

$$\boldsymbol{z} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \omega \\ i \end{bmatrix} + \boldsymbol{v}, \tag{11}$$

where $\boldsymbol{z}$ is the measurement variable, and $\boldsymbol{v}$ is the zero-mean multivariate Gaussian noise. The measurement data were gathered at 0.005 s time intervals.

The models were used for online estimation and diagnosis. In this study, an EKF algorithm was employed. In the estimation stage, the state variable is augmented by the unknown health parameter $\boldsymbol{h}$ and denoted by $\mathbf{x} = \left[\boldsymbol{x}^T, \boldsymbol{h}^T\right]^T$. The state and measurement models in (10) and (11) can be rewritten in recursive matrix form:

State model: $\mathbf{x}_t = F(\mathbf{x}_{t-1}) + \boldsymbol{w}_t$ or

$$\begin{bmatrix} \omega_t \\ i_t \\ k_{T,t} \\ T_{L,t} \end{bmatrix} = \begin{bmatrix} (1 - b \cdot dt/J)\omega_t + dt/J\left(k_{T,t-1} \cdot i_{t-1} - T_{L,t-1}\right) \\ -dt/L \cdot k_{T,t-1} \cdot \omega_{t-1} + \left(1 - \frac{R}{L}\right) i_{t-1} + V \cdot dt/L \\ k_{T,t-1} \\ T_{L,t-1} \end{bmatrix} + \boldsymbol{w}_t. \tag{12}$$

Measurement model: $\boldsymbol{z}_t = H(\mathbf{x}_t) + \boldsymbol{v}_t$ or

$$\boldsymbol{z_t} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \omega_t \\ i_t \\ k_{T,t} \\ T_{L,t} \end{bmatrix} + \boldsymbol{v}_t, \tag{13}$$

where $k_{T,t}$ and $T_{L,t}$ denote the k_T and T_L at the current time t, respectively. Process error $\boldsymbol{w}_t$ is given by the zero-mean multivariate Gaussian noise with covariance, whose diagonal elements are $I \times 10^{-9}$, where I is the identity matrix. The measurement noise $\boldsymbol{v}_t$ is set as 0.1 and 0.01, which can be determined by evaluating the dispersion of the measured data. More details on the EKF can be found in literature [17, 20].

The initial state variables are given as $\mathbf{x}_0 = [0, 0, 1.28, 0.05]^T$ based on the evaluation of k_T and T_L at the initial stage of motor operation. The estimation stage was performed in two steps, as shown in Fig. 1. The first is prior estimation by the state model under a given input $\boldsymbol{u}$. Next is the posterior update by the measurement model, which leads to the estimated state variable and health parameters. Figure 6a, b show the estimated state variable $\boldsymbol{x}$ (velocity and current) and health parameters $\boldsymbol{h}$ (k_T and T_L) along with 95% prediction and confidence intervals, respectively. As shown in Fig. 6b, the health parameters rapidly converged to their true values. The values at the end of the voltage application (i.e., 1.5 s) were used as the estimated health of each component to assess the system performance given by (9).

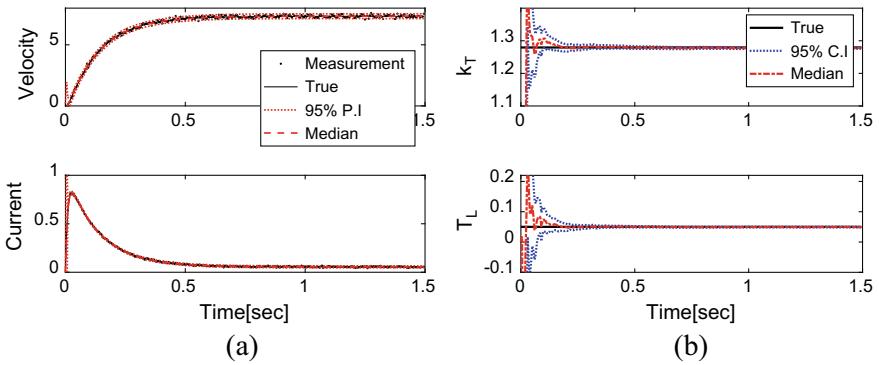


Fig. 6 Online estimation and diagnosis in a cycle by physics-based approach: **a** state variable $\boldsymbol{x}$ and **b** health parameters h

Once the health parameters $\boldsymbol{h}$ are estimated for each cycle, the next step is to transfer them to the offline stage, as shown in Fig. 1. The results are shown in Fig. 7a, c, and e for all three cases; the black dots denote health parameters estimated up to the current cycle. Using these data, the degradation models of each health parameter were fitted, i.e., the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (7) and (8) were estimated. For this purpose, the Markov Chain Monte Carlo (MCMC) method, which determines the parameters by large samples (10^4 in this study) to represent the uncertainty, is applied for the likelihood between the data and the model. Future degradation behaviors were also predicted by extrapolating the model. In Fig. 7a, c, and e, these are represented by the median and 95% predictive interval (PI) curves. Note that the associated uncertainty is so small that it nearly overlaps in this example. Once the health parameters $\boldsymbol{h}$ are predicted for future cycles, they are transferred to the online stage, as shown in Fig. 1. In this case, they are used for the prediction stage wherein only the state variables $\boldsymbol{x}$ are estimated by the state model, because $\boldsymbol{h}$ is known. Subsequently, the system performance in the future cycles obtained as samples are transferred to the offline prognosis over cycles. The results are given by the median and 95% PI in Fig. 7b, d, and f for the three cases, respectively. It is worth noting that establishing a degradation model for the system performance is not necessary because these are obtained from the online estimation stage as samples.

Because true solutions are available, they are superimposed by solid black lines and compared with the predictions. The prediction results at the current cycle agree well with the true solutions for all three cases. It should be noted that the system failure is defined as 70% of its initial value, depicted by the horizontal green line in the figure. The end of life (EOL) was found at 683, 617, and 369 cycles in terms of the median for the three cases. The reason for the shorter life in Case 3 is attributed to the acceleration effect caused by the simultaneous degradation of both components. In Fig. 7a, c, and e, the blue dotted horizontal lines indicate the failure thresholds of each health parameter. They are defined by the corresponding values at EOL when

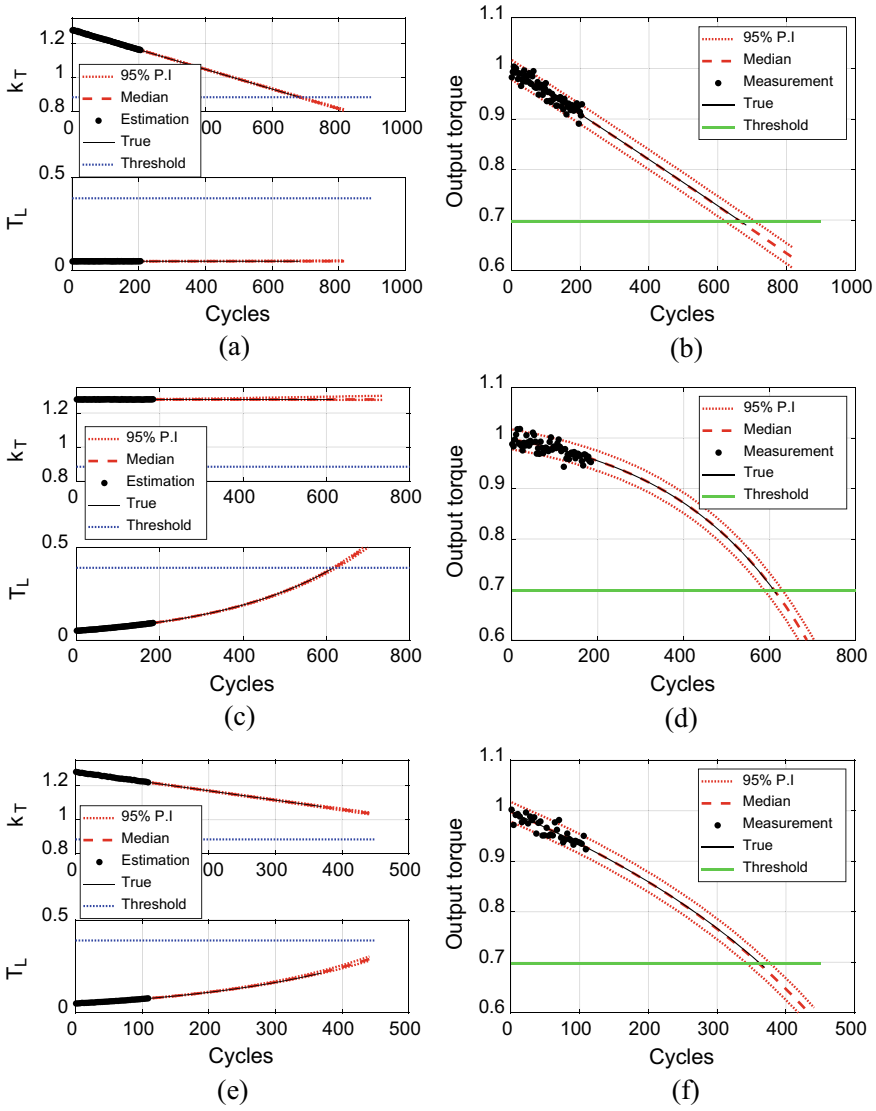


Fig. 7 Offline trending and prognosis over cycles by physics-based approach: **a** components health degradation for Case 1, **b** system health degradation for Case 1, **c** components health degradation for Case 2, **d** system health degradation for Case 2, **e** components health degradation for Case 3 and **f** system health degradation for Case 3

the system undergoes degradation of each component. The threshold values for k_T and T_L are 0.884 and 0.3897 at EOLs 683 and 617 from cases 1 and 2, respectively.

Once the system prognostics information is available, the maintenance effect of individual components on the system health can be evaluated. This process is illustrated in Case 3 wherein two components degrade simultaneously. Note that all subsequent computations are based on median values, unless stated otherwise. First, the current conditions are assessed by introducing the health index (HI) and RUL. The HI indicates the current health status, whereas the RUL estimates how many cycles remain until final failure. HI is defined by the ratio of degradation at the current cycle to that at the EOL and ranges from 0 (normal) to 1 (failure). Because the k_T values at the initial, current, and EOL cycles are 1.28, 1.1, and 0.88, respectively (Fig. 7a), the index is $(1.28 - 1.10)/(1.28 - 0.88) = 0.45$. The indices for the other parameters can be obtained in a similar manner. RUL is defined as the difference between EOL and the current cycle. For k_T and T_L in Fig. 7e, the EOLs are found at the cycles crossing the threshold lines (not shown here), which are approximately 721 and 513, respectively. Because the current cycle is 109, the RULs of k_T and T_L are approximately 609 and 412, respectively, and the system RUL is $361 - 109 = 252$ cycles.

All the results are summarized in Table 2 and plotted using bar charts in Fig. 8a, b. Note in Table 2 that the HI of k_T has degraded (increased) to 0.1507, which is greater (worse) than the 0.0789 of T_L. However, its RUL is approximately 609, which is longer than the 412 of T_L. This is because of their different degrees of influence on the overall system performance.

Subsequently, a what-if study was performed for scenarios in which one of the components was repaired or replaced by a new one. The health parameter is reset to the original value and the system performance is predicted under the renewed condition when the components that influence the health are replaced. These results are shown in Fig. 8c, d when the components that influence factors k_T and T_L are repaired. The new EOLs of the components that influence factors k_T and T_L got extended to approximately 512 and 701 cycles, respectively, thereby yielding new RULs of approximately 403 and 592, respectively, as shown in Fig. 8d. Among the choices regarding which component to repair, repairing the bearing (T_L) is more

Table 2 Results for simulation

Name	Flux	Bearing	System performance
Symbol	k_T	T_L	S
Current cycle	109	109	109
End of life	718.0811	521.1807	361
Remaining useful life	609.0811	412.1807	252
Initial value	1.2796	0.05	0.9971
Current value	1.22	0.0768	0.9278
Threshold	0.884	0.3897	0.6978

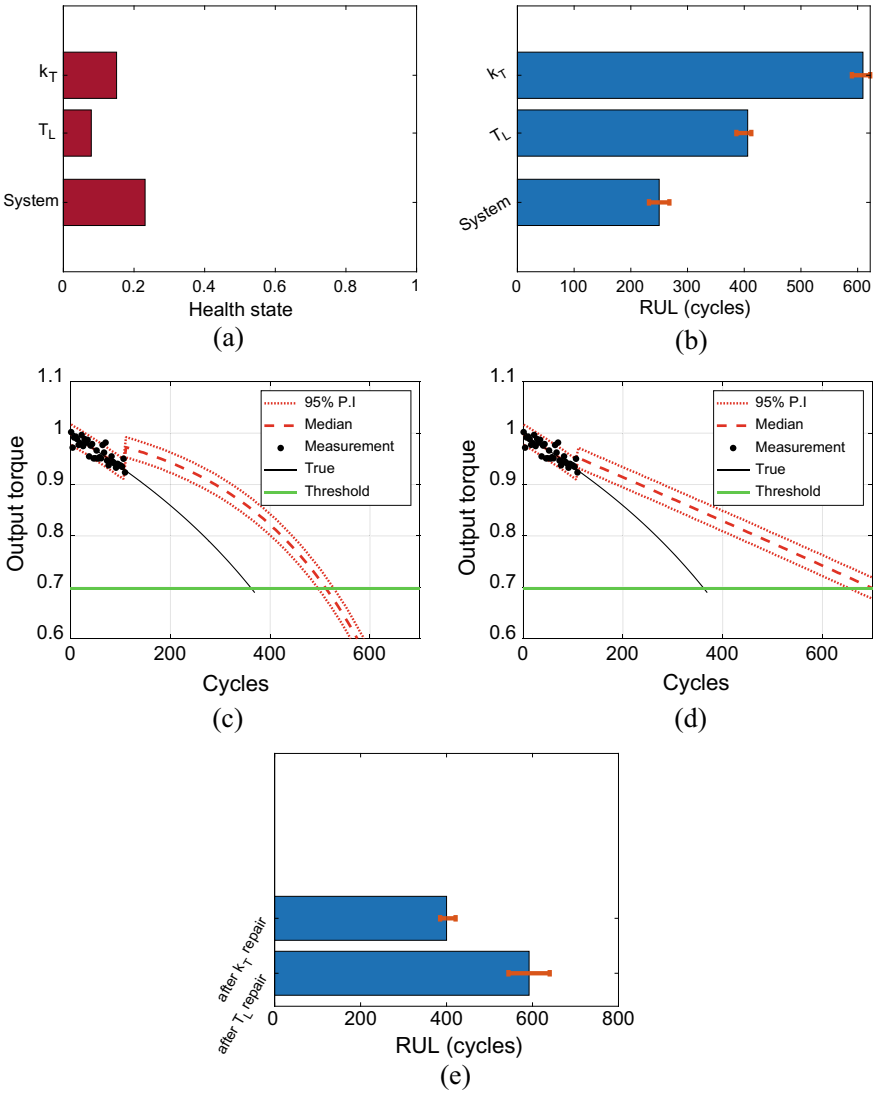


Fig. 8 Maintenance scenario in physics-based approach: **a** current health status, **b** RUL of component and system, **c** prognosis with k_T repair, **d** prognosis with T_L repair and **e** system RUL when component is repaired

desirable as it leads to a longer RUL. In Fig. 8b, e, the red error bar indicates the 95% PI of the RUL prediction. Answers to the following significant questions could be found from this study: what is the current health condition of the components and system, how much longer can the system operate until failure, which component

should be replaced to extend the system life, and how much longer can the system operate after repair.

In the following section, two real-world examples are addressed to illustrate the applications of the proposed framework: motors for the attitude control of satellites and driving quadcopters.

4 Case Study 1: Reaction Wheel Motor in Satellites

4.1 Problem Definition

In this section, reaction wheel motors for the attitude control of satellites are considered. Note that the most of the contents in this case study are from the authors' previous paper [21]. Satellites in space require accurate attitude control and high reliability to conduct their missions fully. The RW actuated by a motor provides consistent angular momentum to help stabilize a satellite against external torsion, such as solar radiation pressure, and to control its precise attitude. However, owing to continuous operation, the functioning of the motor becomes degraded over time, thereby jeopardizing the reliability of the entire satellite control system [22]. According to a survey of the failure statistics of satellite components, most failures are attributed to the actuators of the attitude and control system (AOCS), such as the RW motor [23]. Therefore, the proposed method was implemented to monitor the health and predict the degradation of the RW motor to improve their reliability.

Few studies have addressed the prognostics of RW motors [24, 25]; however, motor RUL prediction is conducted at the single-component level. Motivated by the aforementioned limitations and requirements, this case study addresses the RUL prediction of an RW motor based on the proposed system-level prognostics framework. In this study, we conducted an accelerated life test (ALT) on an RW motor for a period of 3 years to acquire real measurement data with a low sampling rate, similar to a space environment. A proper failure threshold was imposed on the motor based on the characteristic curve given by the design requirement. The RUL is predicted using the degradation relation between the system and its components, assuming that the data are obtained during space operations.

4.2 Experimental Setup

The RW in this study was developed for the Korean Space Launch Vehicle, named the Science and Technology Satellite-3 (STSAT-3), and is addressed in Ref. [26]. It is actuated by a motor to provide consistent angular momentum and control its precise attitude. ALT was performed for this motor; one operation cycle comprises a short-term pull-up followed by a longer period at constant speed. During ALT, the

current and angular velocity signals were acquired at a sampling rate of 2 Hz. The pull-up test lasts only for a few seconds, while the rest of the time is given for the constant-speed test which extends from 10 to 20 h. Consequently, a day is spent for a single-cycle operation, on average, and the entire test lasts for three years. The pull-up operation evaluates motor performance by applying maximum voltage to the motor. The test was conducted under two extreme temperature conditions, i.e., hot (60 °C) and cold (−30 °C), within a thermal vacuum chamber to evaluate its reliability and performance. Even after three years there were no failures, but the test was stopped considering safety and the abnormally high current consumption.

4.3 Application of the Systems Approach

4.3.1 Online Parameter Estimation

In online diagnosis, the EKF is used to estimate the health status based on the motor dynamic model and measured signals from each cycle. The governing equations for the mechanical and electrical parts of the motor are the same as those in Sect. 2. Table 3 lists the model parameters used in this case study. The health parameters responsible for the motor performance degradation are given by $\boldsymbol{h} = [k_T, b]^T$ which are the back EMF and friction coefficients, relating to the permanent magnet health and the bearing condition, respectively. The input operation parameter is the voltage V at current time t.

However, unlike the simulation studies, it is difficult to assign initial values to the process. Measurement of noise covariance and improper values significantly affect and degrade the performance. To overcome this, an Adaptive EKF (AEKF) was employed to adaptively estimate the covariance matrices at each step of the EKF [27]. The forgetting factor ($\alpha = 0.8$) is used for adaptive estimation. Note that a larger α indicates more weight on previous estimates and incurs less fluctuation in the covariance, as well as longer time delays to adapt to changes. In this study, $\alpha = 0.8$ for all studies.

Table 3 Parameter description and values for extended Kalman filter (EKF)

Symbol	Description	Value
R	Armature resistance	22 Ω
L	Armature inductance	0.1215 H
J	Moment of inertia	0.001143 kg m^2
b	Friction coefficient	1.01×10^{-5} N m s/rad
k_T	Electromechanical coupling coefficient	0.054 Nm/A
V	Input voltage	24 V
T_L	Load torque	0.0001 Nm

4.3.2 Motor System Performance

To evaluate the performance of the RW motor, a typical characteristic curve that is defined by the relation between the output torque ($T_{output} = J\dot{\omega}$) and the angular velocity ω during the pull-up range was applied. To ensure the minimum actuation performance of the RW, the motor must generate at least 5 Nm of output torque at $\omega^* = 314.16$ rad/s, i.e., the motor is considered to have failed when the performance falls below this point. This is a design requirement for the STSAT-3 mission, where a satellite with an inertial moment of 18 kg m^2 needs to maneuver 25° in 40 s. In this context, motor system performance is defined by the following expression:

$$P_{sys} = J\dot{\omega}|_{\omega=\omega^*}, \tag{14}$$

and the failure threshold point is given by 5 Nm.

4.3.3 Offline Prognosis and Monitoring

In the offline monitoring and prognosis, the health parameters $\boldsymbol{h}$, estimated from the online diagnosis for each cycle, were transferred and monitored until the current cycle. An empirical degradation model was introduced to quantitatively describe the health degradation over long-term cycles, in which the model parameters were estimated using the accumulated health parameters. A particle filter (PF) was used in this case study to recursively estimate the probability density function (PDF) of the long-term health status and model parameters in the form of particles [28, 29]. The future trend was predicted by extrapolating each particle to future cycles. As in the EKF, the standard PF also consists of the state transition function f and measurement function h, as follows:

$$x_k = f(x_{k-1}, \beta_k), \tag{15}$$

$$z_k = h(x_k, n_k), \tag{16}$$

where k is the cycle step index, x_k is the estimated health state, β_k is the degradation model parameter, z_k is the measurement data (in this case, the health parameter values obtained using the online estimation), and n_k is the measurement noise. To account for degradation, an empirical exponential function was employed for function f [30–32]:

$$f(x_{k-1}, \beta_k) = exp(\beta_k dt)x_{k-1}. \tag{17}$$

The Gaussian PDF assumes the measurement noise, $n_k \sim N(0, \sigma_k)$, where σ_k is the unknown standard deviation. Consequently, the unknown parameter to be estimated is $\theta = [x, \beta, \sigma]^T$. It should be emphasized that the health parameters are

estimated by the motor dynamic model using the AEKF for the online stage, whereas its trend over long-term cycles is estimated by the degradation model using the PF for the offline stage. Once the degradation model is estimated up to the current cycle, it is used to predict future RUL values.

In an offline prognosis, it is often the case that the degradation trend accelerates after a certain cycle or initial fault. To account for this in the PF process, a shifting kernel PF (SKPF) that can detect when the current cycle deviates from the normal is used [33]. To this end, SKPF calculates the likelihood L, and subsequently calculates the decision function d_k:

$$d_k = -\ln\left(\frac{1}{N}\sum_{i=1}^{N} L\left(z_k|x_k^i, \beta_k^i, \sigma_k^i\right)\right). \tag{18}$$

When the observed degradation is close to the normal condition, the likelihood tends to be high and assumes a negative value; therefore, degradation is not monitored. Conversely, when the state degrades in a different fashion, e.g., deviates from the normal, the likelihood becomes lower and assumes a positive value. By monitoring these cycles and examining when the decision function reaches a positive value, the anomaly point is identified. Once detected, the SKPF shifts the kernel function used in the resampling step of the PF and adapts to the new degradation trend.

4.4 Application Results

In this section, the results of applying the AEKF to the system framework are discussed. In the AEKF, the initial values are necessary; they are given as $\boldsymbol{x}_0 = \left[0, 0, 0.054, 10^{-5}\right]^T$ based on the motor specification, and the first two are the state variables $[\omega, i]^T$, whereas the remainder are the health parameters $[k_T, b]^T$, respectively. The initial process and measurement noise covariances were arbitrarily assumed as $Q = \left[10^{-5}\,0; 0\;10^{-5}\right]$ and $R = [2\;0; 0\;0.1]$, respectively. The value at the end of time is then used as the health value of the cycle.

The ALT test ended at 658 and 600 cycles under cold and hot conditions, respectively. Among the results, the degradation of the friction coefficient (b) under cold conditions was noticeable, while the others did not change significantly. Therefore, the test data under cold conditions were used in this study to verify the proposed methodology. Bearing degradation was found to be dominant in this test and was responsible for the motor performance degradation. This conforms to the literature, which indicate that the bearing is the most vulnerable in RW motors.

The characteristic curves obtained from the online diagnosis in each cycle are shown in Fig. 9a, where the x-axis represents the angular velocity, and the y-axis represents the output torque. The graph shows that the slope of the curve constantly decreases as the cycle proceeds and approaches the threshold point. Because the motor system performance is defined by the torque at $\omega^* = 314.14$ rad/s, it is marked

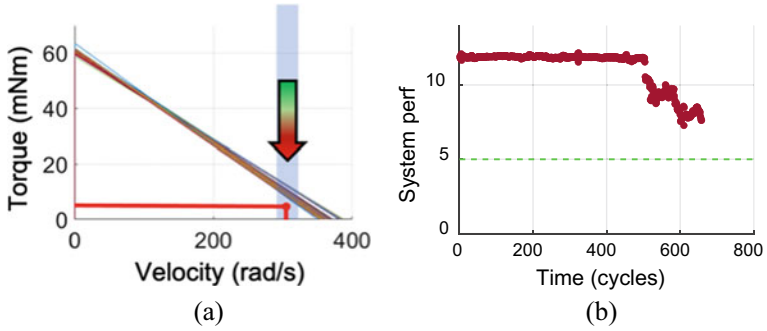


Fig. 9 **a** Characteristic curve and **b** system performance data of cold condition test

by the points in Fig. 9b; the green dashed line indicates the threshold. Even though the test lasted over three years with 658 cycles, the results indicate that the motor did not fail.

Using the data up to 658 cycles, the degradation models of each health parameter were estimated and their future was predicted using the SKPF. The results for both parameters are shown in Fig. 10a, c. The blue dots and triangles represent the estimated health values and anomaly points, respectively, detected by the decision function. The red dashed and solid lines represent the median and 95% PI, respectively. Figure 10b, d represent the trace of the anomaly decision function. The blue line with a circle represents the decision function value, and the red dotted line represents the anomaly threshold set by the user. The results of parameter b, as shown in Fig. 10c, confirm that SKPF successfully detects the initial point of the degradation trend change after 500 cycles. Few anomalies were detected before 500 cycles, which may be attributed to sudden abnormal measurements during normal conditions. When the degradation pattern changed after 500 cycles, the SKPF algorithm successfully adapted to the new degradation trend. The trace of the decision function shows that d_k increases significantly and exceeds the threshold when the estimated state becomes incoherent with the observed data.

On completion of prediction for future cycles, the health parameters are transferred to the online stage and used in the state model to predict the system performance. Subsequently, they are transferred to the offline stage. The results are shown in Fig. 10e with the median and 95% PI. With the system threshold given by the green dotted line, the EOL cycle for the system was predicted to be 808 cycles and the RUL was 150 cycles. It should be noted that the reason for predicting RUL at 658 cycles is that the test ended at this cycle. To validate this prediction, test should be continued further till 808 cycles; however, it was not conducted owing to limited cost and time.

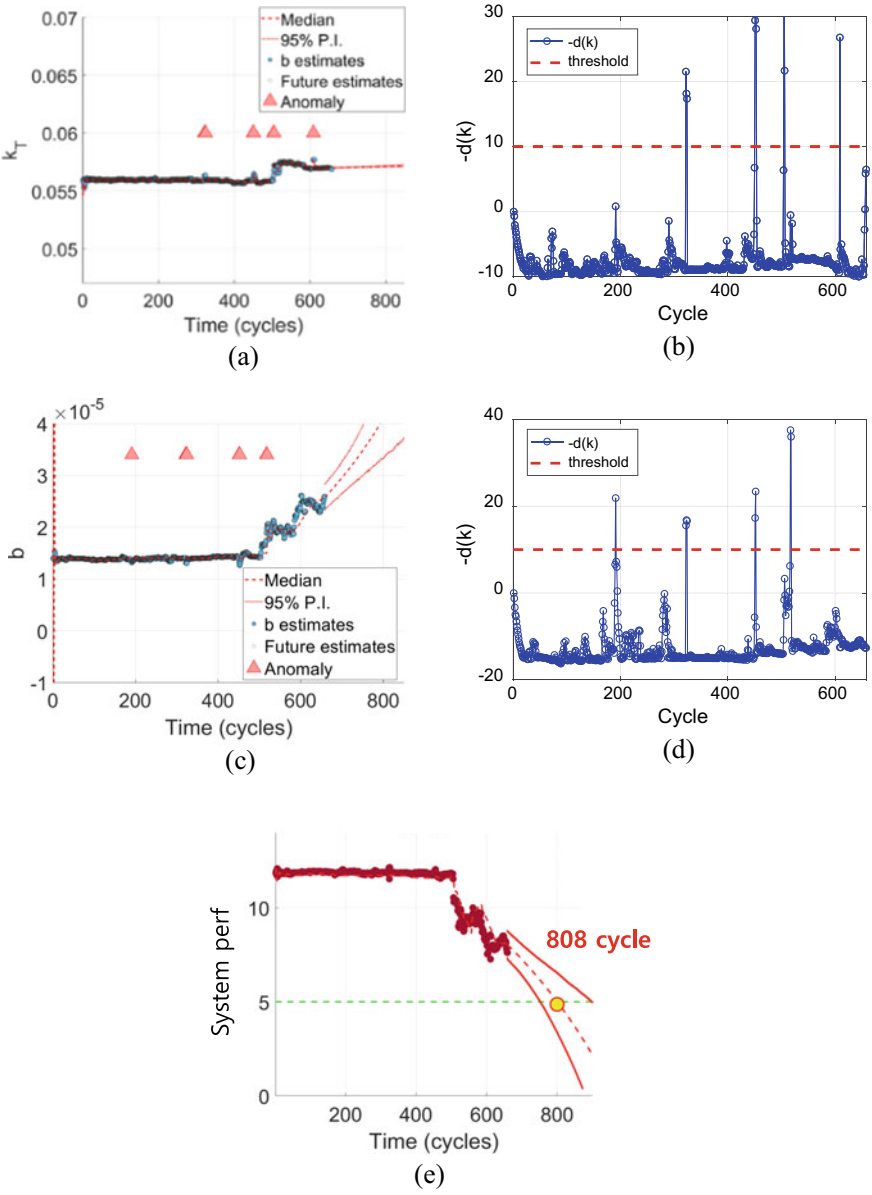


Fig. 10 **a** Prediction of parameter k_T by SKPF algorithm and **b** corresponding decision function **c** prediction of parameter b by SKPF algorithm and **d** corresponding decision function **e** prediction of system performance based on the health parameters

5 Case Study 2: Driving Motors in Quadcopters

5.1 Problem Definition

Since the commercial launch of quadcopters in early 2000s as unmanned aerial vehicles, their use, particularly in the field of aerial imaging, has been burgeoning [34]. Several commercial services related to agriculture activities, traffic control, and delivery of goods have been launched for quadcopters since the mid-2010s; however, safety issues owing to quadcopter failure have become a concern, with a high-risk failure factor being the occurrence of falls owing to the performance degradation of driving motors [35].

Based on a literature survey, it was found that most studies on the health diagnosis of quadcopter motors have been conducted to aid the design of robust flight controllers [36–38]. Therefore, they do not focus on health management or failure prevention based on the PHM framework. The recent studies in this direction that exploit parameters such as the vibration, current, or rotational speed have mainly been sensor-based approaches requiring attachment of additional sensors to the quadcopter [39–42].

The quadcopter can record various flight information, such as posture and angular velocity, position and linear velocity, and motor control data during flight. By exploiting these data and KF algorithms, it is possible to estimate the forces and moments acting on the aircraft and the degradation of the motor performance. In this case study, a PHM framework is presented that evaluates the health of individual motors and predicts their RUL based on the systems approach. It is applied to a Parrot Mambo drone (PMD), a micro quadcopter, to demonstrate the RUL prediction of the driving motors. Practically, the PMD structure is vulnerable to failure, and placing sensors to diagnose motor conditions is challenging. The PMD is more suitable for application of the framework proposed in this study.

5.2 Experimental Setup

The PMD used in this case study is a miniature quadcopter manufactured and sold by Parrot, France, measuring 7.1 $\times$ 7.1 inches and weighing 63 g [43]. The parameters for the PMD quadcopter dynamic analysis that were obtained using the MathWorks Simulink Parrot Minidrone model and actual measurements are summarized [44, 45]. The PMD is equipped with an 8520 coreless DC motor whose parameters were obtained from experiments based on the data from previous studies [46].

The overall framework for prognostics of quadcopter motor comprises two phases as shown in Fig. 11. The first step in online diagnosis is the state estimation of the quadcopter using a KF. The force and moment acting on the aircraft, as well as the rotational speed of the four motors were estimated using the flight data in the dynamics model of the quadcopter. The second step in the online diagnosis is the

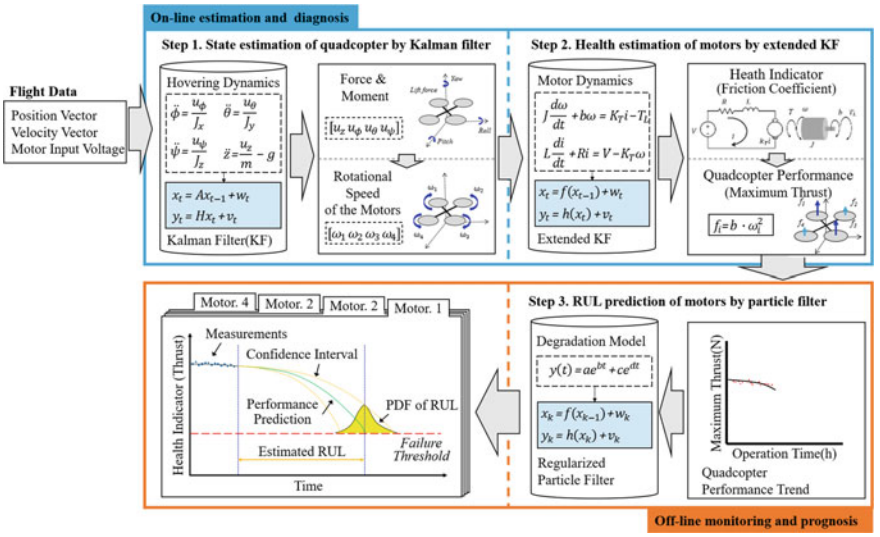


Fig. 11 The Quadcopter PHM framework

health estimation of motors using an EKF. The friction coefficient representing the degree of motor degradation was estimated by applying the estimated rotational speed of each motor to its dynamic model. Consequently, thrust, which represents a HI of the motor, was obtained. Offline observations identify the decreasing trend of thrust in each motor over long-term cycles and predict the RUL until failure. To realize this, an empirical degradation model was introduced, and the RUL of each motor was predicted using a regularized PF (RPF). Here, failure is defined as a situation in which the quadcopter cannot hover, i.e., it cannot sustain its own weight during lift-off.

5.3 Application of the Systems Approach

5.3.1 State Estimation of Quadcopter Using KF

The first step in online diagnosis involves the application of flight data collected during hovering to the quadcopter dynamics model to estimate the force and moment applied to the aircraft. The rotational speeds generated by the motors are estimated based on data obtained from hovering flights in this study. During hovering, there are nearly no roll, pitch, or yaw motions to maintain posture. Consequently, the equations for the translational and rotational degrees of freedom are obtained as follows:

$$\ddot{z} = \frac{u_z}{m} - g, \ddot{\phi} = \frac{u_\phi}{J_x}, \ddot{\theta} = \frac{u_\theta}{J_y}, \ddot{\psi} = \frac{u_\psi}{J_z} \tag{19}$$

In this equation, the vector $[z\ \phi\ \theta\ \psi]^T$ denotes the altitude z and rotational angles, i.e., the roll, pitch, and yaw motions, respectively, and its 2nd derivative $\left[\ddot{z}\ \ddot{\phi}\ \ddot{\theta}\ \ddot{\psi},\right]^T$ is the vertical acceleration and corresponding rotational-angular accelerations, respectively. The vector $\left[m\ J_x\ J_y\ J_z\right]^T$ is the mass and rotational inertia, and g is the gravitational acceleration. The vector $\left[u_z\ u_\phi\ u_\theta\ u_\psi\right]^T$ is the vertical force and moment acting on the quadcopter aircraft.

The flight data collected during hovering are the position vector $[z\phi\ \theta\ \psi]^T$ and its derivative: the velocity vector $\left[\dot{z}\ \dot{\phi}\ \dot{\theta}\ \dot{\psi},\right]^T$. By applying these to (19), the vector of the vertical force and moments $\left[u_z\ u_\phi\ u_\theta\ u_\psi\right]^T$ can be estimated. To implement this, a KF in which the system and measurement models are defined in recursive form, is applied [47]. Once the force and moments are obtained, the rotational speed ω_i of each motor can be obtained as follows [48].

$$\begin{bmatrix} \omega_1^2 \\ \omega_2^2 \\ \omega_3^2 \\ \omega_4^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4c} & \frac{-1}{4cl} & \frac{-1}{4cl} & \frac{-1}{4d} \\ \frac{1}{4c} & \frac{-1}{4cl} & \frac{1}{4cl} & \frac{1}{4d} \\ \frac{1}{4c} & \frac{1}{4cl} & \frac{1}{4cl} & \frac{-1}{4d} \\ \frac{1}{4c} & \frac{1}{4cl} & \frac{-1}{4cl} & \frac{1}{4d} \end{bmatrix} \begin{bmatrix} u_z \\ u_\phi \\ u_\theta \\ u_\psi \end{bmatrix} \tag{20}$$

where ω_i are the rotational speed of each motor, respectively; c is the motor thrust coefficient; l is the length of the arm; and d is the motor drag coefficient. The motor thrust f_i is proportional to the square of the speed ω_i^2 and is expressed as follows:

$$f_\mathrm{i} = c \cdot \omega_\mathrm{i}^2 \tag{21}$$

5.3.2 Health Estimation of Motors by Extended KF

In the second step of the online diagnosis, the friction coefficient of each motor was estimated by applying the estimated rotation speed of the motor to its dynamic model. The motor dynamic model comprises the same governing equations as the physics-based approach described in Sect. 2. When a cycle lasts for an extended period, performance degradation of the motor occurs due to various factors. The most representative is an increase of frictional force owing to wear of mechanical parts, such as bearings or brushes, which corresponds to coefficient b [49].

Once the motor friction coefficient b is estimated using the EKF, it can be directly used as a health indicator. However, it is preferable to use the maximum thrust under the corresponding degraded condition. This is because the quadcopter fails when the sum of the maximum thrust of the four motors is lower than the thrust required to maintain the takeoff and hovering of the quadcopter. By exploiting this in the prognosis, RUL can be predicted using this as a failure threshold. The threshold can be calculated using the following formula:

$$F_{hov} = m(g + a) + f_d,\ f_d = C_d \frac{1}{2} \rho v^2 A \tag{22}$$

where F_{hov} is the thrust required for hovering, m is the mass of the quadcopter, g is the acceleration due to gravity, a is the takeoff acceleration, f_d is the drag force, C_d is the drag coefficient, ρ is the air density, v is the takeoff speed, and A is the cross-sectional area of the quadcopter in the horizontal plane [44]. The maximum thrust of the motor is obtained by calculating the rotation speed using the motor dynamic model under the current value of the friction coefficient at the maximum input voltage condition and applying it to (21).

5.3.3 RUL Prediction of Motors by PF

The maximum thrust of the motor decreases with each cycle as the motor performance degrades due to the increase in mechanical friction. During offline monitoring, these cycle trends are monitored in two stages. First, the anomaly detection which detects the cycle at which a deviation occurs from the normal condition due to fault development. Next, the degradation prediction after anomaly detection in which the thrust begins to decrease exponentially as the cycle continues. The degradation trend is suitably described by introducing an empirical model and RUL until failure is predicted.

In this study, the anomaly is detected by the Naïve Bayes classifier which explores the first prediction time (FPT), i.e., the cycle point where the normal and fault conditions are divided [50]. Once the anomaly is detected, it is presumed that the degradation begins to increase, and an empirical degradation model is employed to describe this trend. The PF algorithm is used to estimate the model parameters and predict the RUL, similar to that of *Case study 1*.

5.4 Application Results

5.4.1 Online Estimation and Diagnosis of the PMD Motors

As the first step of online diagnosis, the KF is used to estimate the rotational speed of each motor during 5–50 s of hovering motion. In the KF, the standard deviations of the process and measurement noises are given by 0.1 and 1×10^{-5}, respectively. The mean values are 1809.12, 1809.12, 1809.12, and 1809.11 rad/s, and the rotational speeds are nearly identical to maintain a stationary posture during hovering. To calculate the maximum thrust as the second step of the online diagnosis, the motor friction coefficient was first estimated using the dynamic model of the motor and EKF. The results are shown in Fig. 12. The standard deviations of the process and measurement noises were 1×10^{-13} and 1×10^{-3}, respectively. As is evident from these figures, the friction coefficient rapidly converges to a constant value, although it

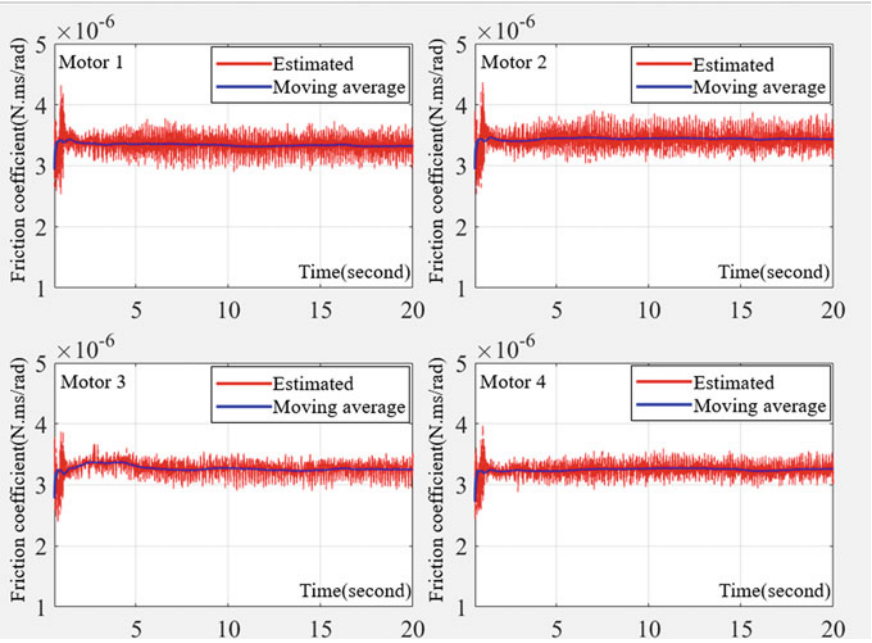


Fig. 12 Friction coefficient of motors during PMD hovering

tends to fluctuate around the mean, which is 3.31×10^{-6}, 3.43×10^{-6}, 3.27×10^{-6}, and 3.25×10^{-6}, respectively. The maximum thrust of each motor can be obtained by applying the maximum input voltage, which was 2.7 V in this case. After passing through the motor dynamic model under this condition and the friction coefficient of each motor, the rotational speeds were obtained as 2217, 2199, 2224, and 2226 rad/s. Consequently, the maximum thrust becomes 0.232, 0.2282, 0.2335, and 0.2339 N, respectively, according to (21).

5.4.2 Offline Monitoring and Prognosis

To implement the RUL prediction of the PMD motors, Motor 4 was chosen to perform accelerated degradation until failure occurred, which occurred after 106 h. During degradation, 48 cycles of hovering tests were performed at intermittent intervals and flight data were collected. The target was that the altitude should be maintained at 1.1 m with a rotational angle of 0 rad during the hovering mode. The RUL of Motor 4 was predicted using the RPF. The failure threshold of the PMD was determined as $0.2224N$ using (22). In the calculation, a is 2.5 m/s^2, f_d is 0.0776 N, C_d is 0.0624, ρ is 1.225 kg/m^3, v is 2.5m/s, and A is 0.325m^3.

The maximum thrust was obtained by the KF-based online estimation at each cycle for the motors until the 48th cycle. Figure 13 shows the results for Motor 1

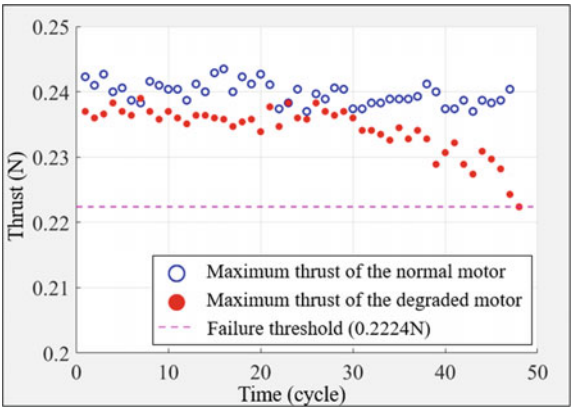


Fig. 13 Online estimation of maximum thrust at each cycle

(normal) and Motor 4 (degraded), and it can be observed that Motor 1 maintains a maximum thrust value between 0.2435 and 0.237 N, whereas Motor 4 gradually degrades after approximately 30 cycles.

Two exponential functions were employed for the RPF degradation model as follows.

$$f = \beta_1 exp(\beta_2 t) + \beta_3 exp(\beta_4 t) \tag{23}$$

where f is the maximum thrust of the motor, t is the long-term cycle index (don't be confused with the time in on-line diagnosis), and $\beta_i (i = 1, \ldots, 4)$ are the parameters in the degradation model.

$$x_0 \sim U(0.23, 0.25), \beta_2 \sim U(0.10, 0.15), \beta_3 \sim U(0.2, 0.3),$$
$$\beta_4 \sim U(-0.1, 0.0), and\ \sigma \sim U(0, 0.01).$$

From these, 3000 particles were generated for use in the subsequent process, and the thrust at the current cycle as well as the future values were predicted recursively using the RPF. Figure 14 shows the RUL prediction performed at 39 cycles. The filled and empty black dots represent the data collected up to the current and future cycles until failure, respectively. The red dashed and dotted lines after 39 cycles illustrate the median and 90% PI of the future thrust prediction. The magenta-colored horizontal line represents the failure threshold, and the black vertical line represents the anomaly point. The prediction results show excellent performance because the true EOL resides within the PI and is close to the median of the predicted distribution.

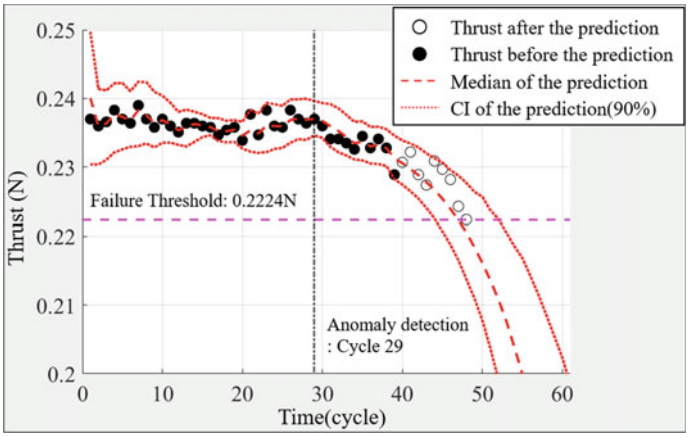


Fig. 14 RUL prediction at cycle 39

6 Conclusion

In this study, a system framework that takes advantage of the motor dynamics and KF estimation was presented to conduct the PHM of DC motors using a model-based approach. Most previous studies treat the motor as a single component and used a data-driven approach in which the raw signal of electric current or vibration was extracted and used for RUL prediction. However, in this study, the motor was considered as a system with multiple components; the health of the components was estimated individually, from which the RUL with respect to the motor system performance was predicted. In the literature, the degradation model for the system was typically used to predict future behavior. However, in this study, it was not introduced but obtained because of component degradation. The proposed framework is validated using two case studies, that is, a satellite RW and the driving motors of a quadcopter. The results demonstrate that the proposed method can provide an effective means to aid decision making in practical applications for DC motors.

Several benefits are expected from this approach; however, we do not have to generate an enormous volume of run-to-fail data for training the data-driven models because we employed the model-based approach. The approach can also be applied in a straightforward manner to other types of motors, provided that the associated model parameters are available or measured a priori. This is in contrast to the data-driven approach, which requires training whenever a motor is changed.

Acknowledgements This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A4A4079904) and the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under the Fostering Global Talents for Innovative Growth Program (P0017307) supervised by the Korea Institute for Advancement of Technology (KIAT). In addition, we would like to thank Editage (www.editage.co.kr) for English language editing.

References

1. Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D (2014) Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. Mech Syst Signal Process 42(1–2):314–334
2. Sun B, Zeng S, Kang R, Pecht MG (2012) Benefits and challenges of system prognostics. IEEE Trans Reliab 61(2):323–335
3. Lessmeier C, Kimotho JK, Zimmer D, Sextro W (2016) Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark data set for data-driven classification. In: PHM Society European conference, vol 3, No 1
4. Li N, Zhou R, Hu Q, Liu X (2012) Mechanical fault diagnosis based on redundant second generation wavelet packet transform, neighborhood rough set and support vector machine. Mech Syst Signal Process 28:608–621
5. Kliman GB, Koegl RA, Stein J, Endicott RD, Madden AM (1988) Noninvasive detection of broken rotor bars in operating induction motors. IEEE Trans Energy Convers 3(4):873–879
6. Grubic S, Aller JM, Lu B, Habetler TG (2008) A survey on testing and monitoring methods for stator insulation systems of low-voltage induction machines focusing on turn insulation problems. IEEE Trans Industr Electron 55(12):4127–4136
7. Nandi S, Toliyat HA, Li X (2005) Condition monitoring and fault diagnosis of electrical motors—A review. IEEE Trans Energy Convers 20(4):719–729
8. Su H, Chong KT (2007) Induction machine condition monitoring using neural network modeling. IEEE Trans Ind Electron 54(1):241–249
9. Zidani F, Benbouzid MEH, Diallo D, Naït-Saïd MS (2003) Induction motor stator faults diagnosis by a current Concordia pattern-based fuzzy decision system. IEEE Trans Energy Convers 18(4):469–475
10. Martínez-Morales JD, Palacios E, Campos-Delgado DU (2010) Data fusion for multiple mechanical fault diagnosis in induction motors at variable operating conditions. In: 2010 7th International conference on electrical engineering computing science and automatic control. IEEE, pp 176–181
11. Wen L, Li X, Gao L, Zhang Y (2017) A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans Industr Electron 65(7):5990–5998
12. Gangsar P, Tiwari R (2020) Signal based condition monitoring techniques for fault detection and diagnosis of induction motors: a state-of-the-art review. Mech Syst Signal Process 144:106908
13. Kim S, Kim NH, Choi JH (2021) A study toward appropriate architecture of system-level prognostics: physics-based and data-driven approaches. IEEE Access 9:157960–157972
14. Kim S, Choi JH, Kim NH (2021) Challenges and opportunities of system-level prognostics. Sensors 21(22):7655
15. Rahimi A, Kumar KD, Alighanbari H (2017) Fault estimation of satellite reaction wheels using covariance based adaptive unscented Kalman filter. Acta Astronaut 134:159–169
16. El Sayed W, Abd El Geliel M, Lotfy A (2020) Fault diagnosis of PMSG stator inter-turn fault using extended Kalman filter and unscented Kalman filter. Energies 13(11):2972
17. Singleton RK, Strangas EG, Aviyente S (2014) Extended Kalman filtering for remaining-useful-life estimation of bearings. IEEE Trans Industr Electron 62(3):1781–1790
18. Zhu X, Zhang H, Xi J, Wang J, Fang Z (2015) Robust speed synchronization control for clutchless automated manual transmission systems in electric vehicles. Proc Inst Mech Eng Part D: J Autom Eng 229(4):424–436
19. Salem T, Haskew TA (1995) Simulation of the brushless DC machine. In: Proceedings of the twenty-seventh southeastern symposium on system theory. IEEE, pp 18–22
20. Bavdekar VA, Deshpande AP, Patwardhan SC (2011) Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter. J Process Control 21(4):585–601

21. Park HJ, Kim S, Lee J, Kim NH, Choi JH (2022) System-level prognostics approach for failure prediction of reaction wheel motor in satellites. Adv Space Res. https://doi.org/10.1016/j.asr.2022.11.028
22. Hu D, Sarosh A, Dong YF (2012) A novel KFCM based fault diagnosis method for unknown faults in satellite reaction wheels. ISA Trans 51(2):309–316
23. Ji XY, Li YZ, Liu GQ, Wang J, Xiang SH, Yang XN, Bi YQ (2019) A brief review of ground and flight failures of Chinese spacecraft. Prog Aerosp Sci 107:19–29
24. Muthusamy V, Kumar KD (2022) Failure prognosis and remaining useful life prediction of control moment gyroscopes onboard satellites. Adv Space Res 69(1):718–726
25. Rahimi A, Kumar KD, Alighanbari H (2020) Failure prognosis for satellite reaction wheels using Kalman filter and particle filter. J Guid Control Dyn 43(3):585–588
26. Kim DH, Yang S, Cheon DI, Lee S, Oh HS (2010) Combined estimation method for inertia properties of STSAT-3. J Mech Sci Technol 24(8):1737–1741
27. Akhlaghi S, Zhou N, Huang Z (2017) Adaptive adjustment of noise covariance in Kalman filter for dynamic state estimation. In: 2017 IEEE power & energy society general meeting. IEEE, pp 1–5
28. An D, Choi JH, Kim NH (2013) Prognostics 101: A tutorial for particle filter-based prognostics algorithm using Matlab. Reliab Eng Syst Saf 115:161–169
29. Orchard ME, Vachtsevanos GJ (2009) A particle-filtering approach for on-line fault diagnosis and failure prognosis. Trans Inst Meas Control 31(3–4):221–246
30. Barbieri F, Hines JW, Sharp M, Venturini M (2015) Sensor-based degradation prediction and prognostics for remaining useful life estimation: validation on experimental data of electric motors. Int J Progn Health Manage 6(3):1–20. https://doi.org/10.36001/ijphm.2015.v6i3.2285
31. Bejaoui I, Bruneo D, Xibilia MG (2020) A data-driven prognostics technique and rul prediction of rotating machines using an exponential degradation model. In: 2020 7th International conference on control, decision and information technologies (CoDIT), vol 1. IEEE, pp 703–708
32. Yang F, Habibullah MS, Shen Y (2021) Remaining useful life prediction of induction motors using nonlinear degradation of health index. Mech Syst Signal Process 148:107183
33. Kim S, Park HJ, Choi JH, Kwon D (2020) A novel prognostics approach using shifting kernel particle filter of Li-ion batteries under state changes. IEEE Trans Ind Electron 68(4):3485–3493
34. Air & Space Magazine (2017) A brief history of quadrotors. Available https://www.smithsonianmag.com/air-space-magazine/brief-history-quadrotors-180963372/
35. Susini A (2015) A technocritical review of drones crash risk probabilistic consequences and its societal acceptance. RIMMA risk information management, risk models, and applications. LNIS 7:27–38
36. Zhong Y, Zhang Y, Zhang W, Zuo J, Zhan H (2018) Robust actuator fault detection and diagnosis for a quadrotor UAV with external disturbances. IEEE Access 6:48169–48180
37. Avram RC, Zhang X, Muse J (2017) Quadrotor actuator fault diagnosis and accommodation using nonlinear adaptive estimators. IEEE Trans Control Syst Technol 25(6):2219–2226
38. Asadi D, Ahmadi K, Nabavi SY (2022) Fault-tolerant trajectory tracking control of a quadcopter in presence of a motor fault. Int J Aeronaut Space Sci 23(1):129–142
39. Zahra N, Buldan RS, Nazaruddin YY, Widyotriatmo A (2021) Predictive maintenance with neural network approach for UAV propulsion systems monitoring. In: 2021 American control conference (ACC). IEEE, pp 2631–2636
40. Pourpanah F, Zhang B, Ma R, Hao Q (2018) Anomaly detection and condition monitoring of UAV motors and propellers. In: 2018 IEEE sensors. IEEE, pp 1–4
41. Kong W, Bian S, Li X, Wang C, Wang J (2021) A new integrated health management for quadrotors based on deep learning. In: 2021 IEEE 10th data driven control and learning systems conference (DDCLS). IEEE, pp 1418–1423
42. Darrah T, Kulkarni CS, Biswas G (2020) The effects of component degradation on system-level prognostics for the electric powertrain system of UAVs. In: AIAA Scitech 2020 forum, p 1626
43. Parrot, Mambo range documentation. Available https://www.parrot.com/en/support/documentation/mambo-range

44. MathWorks, fly a parrot minidrone using hover parrot minidrone simulink template. Available https://kr.mathworks.com/help/supportpkg/parrot/ug/fly-a-parrot-minidrone-using-the-hover-simulink-model.html#mw_7b5d113d-3c79-42eb-838d-c063463e5bb0
45. De Simone MC, Russo S, Ruggiero A (2015) Influence of aerodynamics on quadrotor dynamics. In: Recent researches in mechanical and transportation systems influence, pp 111–118
46. Karnavas YL, Chasiotis ID (2016) PMDC coreless micro-motor parameters estimation through grey wolf optimizer. In: 2016 XXII International conference on electrical machines (ICEM). IEEE, pp 865–870
47. Welch G, Bishop G (2006) An introduction to the Kalman filter. Technical report TR 95:041. Department of Computer Science University of North Carolina. NC, USA
48. Kaplan MR, Eraslan A, Beke A, Kumbasar T (2019) Altitude and position control of parrot mambo minidrone with PID and fuzzy PID controllers. In: 2019 11th international conference on electrical and electronics engineering (ELECO). IEEE, pp 785–789
49. Kulkarni CS, Corbetta M, Robinson EI (2021). Systems health monitoring: integrating FMEA into Bayesian networks. In: 2021 IEEE aerospace conference (50100). IEEE, pp 1–11
50. Wang G, Xiang J (2021) Remain useful life prediction of rolling bearings based on exponential model optimized by gradient method. Measurement 176:109161

Exploratory Fault Detection with Multivariate Data: A Case Study on Engine Bearing

An-Kuo Chao, Min Huang, and Loon Ching Tang

Abstract This paper presents a case study on using statistical method for detecting impending bearing failures using in-situ field data. We first explore the relationships between a few variables of interest using a matrix plot. By focusing on variables with consistent profile, we analyze the change in these multivariate data over time and propose a way to pinpoint impending failure. Due to the way data are generated and the inherent large variation, a Gaussian mixture model (GMM) is proposed and methods analogous to multivariate SPC are then applied to detect "out-of-control" signal. In particular, a phase I analysis using variances corresponding to the within and between sorties variations so that the correct control limits can be determined. From the actual failure and known conditions from field data, it was found that the proposed method is able to signal impending failure before it occurred.

1 Introduction

One of the catastrophic failures of the aircraft engine is the failure of its bearing; particularly during flight [1]. Early detection of bearing failure has been of great importance, particularly for the military and aviation industries. In this paper, we examine a set of in-flight data and identify the most sensitive subset which has demonstrated great promise in flagging impending failures.

Engine bearing operates under extreme stresses as it not only supports the main engine shaft that rotates at high speed but also withstands gyroscopic loading generated from tail-spinning maneuvers. In addition to extreme loadings, the engine bearing also operates under high temperature that results in the thinning of lubricant and, in turn, leads to higher wear rates. It was reported that this has been the major cause of failure [2]. Traditionally, engineers have been relying on oil debris/condition

A.-K. Chao · L. C. Tang (✉)
Department of Industrial Systems Engineering & Management, National University of Singapore, Singapore, Singapore
e-mail: isetlc@nus.edu.sg

M. Huang
School of Reliability and Systems Engineering, Beihang University, Beijing, China

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_22

monitoring, and vibration analysis to provide early warning signals [3]. Much effort has been devoted to spectrum analysis in the bearing characteristic frequency range and enveloping analysis in the high frequency range [4]. Most of these studies utilized data collected in laboratories with specially designed test rigs. However, during field deployment, vibration spectrum changes as the bearings deteriorate over time. Quite often, the prognostic capability is rather weak, giving rise to false alarms that led to unnecessary dismantling of engine assembly and diminishing confidence of the ground crews in such analysis. Here, we present a statistical approach in detecting impending failures of aircraft bearings based on a data set from a fleet of military aircrafts.

While literature in failure prediction is vast, that for engine bearing is scarce. To the best of our knowledge, this work represents the first attempt to provide early warning for engine bearing failure using in-flight data. Unlike data collected in a laboratory, in-flight data are usually very noisy. In particular, unlike commercial aircrafts which typically fly at cruising speed, military aircrafts often maneuver sprightly, and that induces large variations in in-flight data. As a result, it is very challenging to detect impending bearing failure as anomaly is confounded by natural variations with extremely low signal to noise ratio. Operationally, should there be any abnormal signals, the aircraft will then be inspected for potential causes of failure. This will typical involve dismantling of the engine assembly which is a major work. It is thus important that false alarm rate is small so that users are confident of the proposed methodology. At the same time, the proposed method must pick up all impending failures as the consequence of the failure during operations is catastrophic. Fortunately, almost all mechanical failures are results of degradation such that signals which are indicative of the engine condition will exhibit deteriorating trend over time.

In the following, we give a brief description of the real life data used in this paper and present their preliminary analysis. This is followed by the presentation of the GMM and EM algorithm for parameter estimation and model validation. The proposed multivariate analysis through the use of multivariate T2 statistics is then presented with particular emphasis on how the between sorties variance could be estimated for the phase I analysis. The resulting control chart is then applied to the engine data set for detecting impending failures. Finally, the conclusion follows.

2 Exploratory Data Analysis

In this paper, in-flight data were collected from a fleet of military turbofan engines with the typical two-stage (low- and high-pressure) engine design. The key variables consisting of rotating speeds at high and low pressure sections, the acceleration and the velocity of vibration, lubricant oil pressure, lubricant oil temperature, etc. were downloaded from the aircrafts after each sortie (this term refers to a trip that an aircraft makes). These data were recorded every second by sensors installed in these aircrafts during operations (i.e. before takeoff, during flight and after landing). To

maintain confidentiality, data are masked by transformation. In particular, the two rotating speeds are expressed as percentage of their maximum speed.

The durations of sorties depend on the missions and thus are of varying length ranging from 30 min to more than one hour. The number of sorties recorded for these engines were left censored, dating back to the latest 30 to 270 sorties. Among a total of 38 engines, one engine bearing had failed during flight, 12 aircrafts had reported abnormality such as excessive vibration or excessive debris collected at oil filter, and 25 were deemed to be working fine. Our key objective is to propose a statistical method to pick up signals for the failed engine in advance.

Figure 1 depicts the matrix plot of the 4 key variables; the two engine rotating speeds (R1 and R2), the velocity of vibration (V) and lubricant oil pressure (P). There is a clear functional relationship between the two rotating speeds while the correlations between other variables are also obvious and strong. Besides anomaly in individual variables, changes in these correlation structures may also suggest potential problems. Unfortunately, these changes are not always discernible from simple scatter plots. For example, Fig. 2 shows the operating profile of the two rotating speeds for a good engine bearing (includes all sorties) and those from the failed engine bearing at the last five sorties prior to failure over the entire operation period. While R1 and R2 seem to be connected through a logistic function which is pinned down at both ends with variation in between, there is no discernible difference between data from a good engine bearing and data of the failed engine bearing. Cranking the data without some pre-processing or stratification is unlikely to yield fruitful insight to the problem.

Here, we stratify the data into the three distinct phases of flight operations, i.e. before takeoff, during flight, and after landing. As illustrated in Fig. 3, which depicts two rotation speeds of the engine for a single sortie, the speeds for takeoff and landing is at about $(R1, R2) = (30\%, 70\%)$. This is used as the cut-off speeds for segregating the three phases. It can also be seen from Fig. 3 that some manual judgment is needed to make the segregation of data for each sortie as the rotating speeds is neither monotone increasing before takeoff or monotone decreasing after landing.

It is noteworthy that as damages were mostly sustained during flight, analyzing data after landing may be the soonest that damages could be picked up without interrupting the mission and ensuing inspection can be initiated. On the other hand, pre-flight check is necessary to ensure that mission is not compromised by mechanical failure. It is thus important to stratify data in phases to facilitate fault detection and diagnosis.

3 The Statistical Framework

After identifying the three distinct phases and the four key variables, namely, two rotating speeds, vibration velocity and oil pressure, in this section, we present the statistical framework for data analysis.

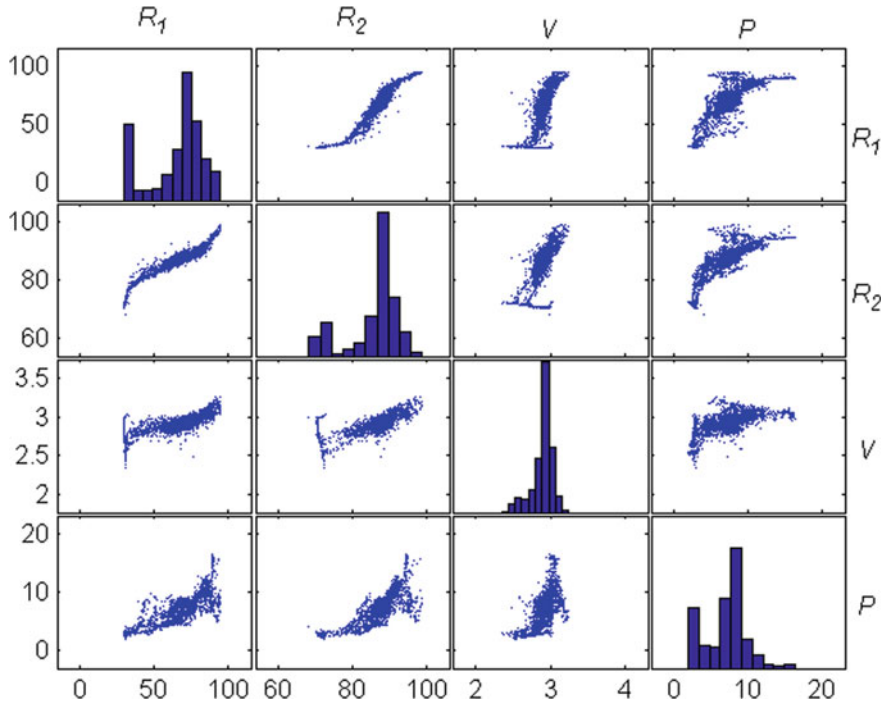


Fig. 1 Matrix plot of the four variables in a typical sortie

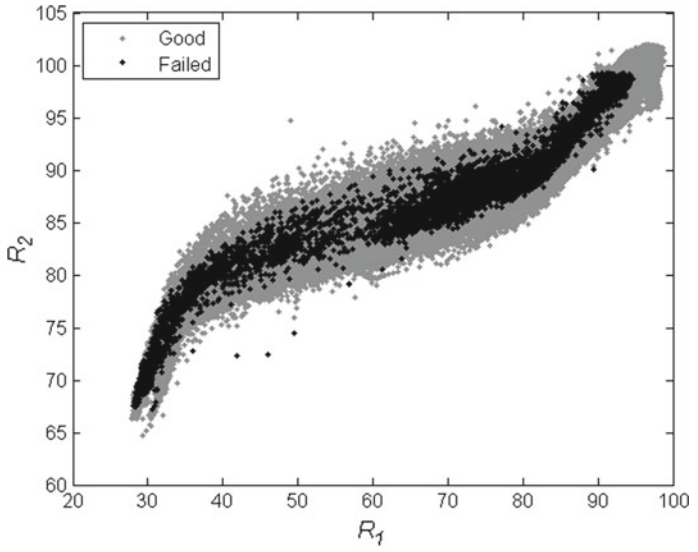


Fig. 2 The scatter plot of the two rotating speeds of good and failed bearings

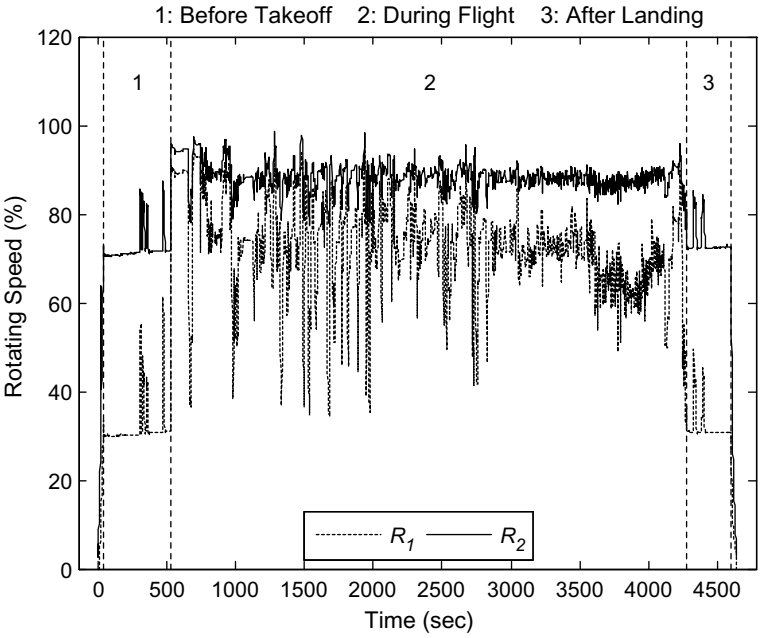


Fig. 3 The two rotating speeds of a typical sortie

Figure 4 displays the four variables of a typical sortie before takeoff and after landing. It can be seen that the two rotating speeds are quite stable except several transitions to higher speeds. As the rotating speeds transited, the velocity of vibration and lubricant oil pressure would also transit. Analysis based on data including these transitions may lead to a biased result. Therefore, statistical classification methods can be applied to separate the transitions from the stable state. Besides, as the rotating speeds are the controllable variables, it is desirable to compare the velocity of vibration and lubricant oil pressure between sorties while keeping the rotating speeds at the stable state. A representative statistic should be computed based on data from the stable state for each variable.

Among various statistical classification methods, the Gaussian mixture model (GMM) has been successfully applied to profile monitoring (see, e.g. [5–7]). For an arbitrary sample by $\mathbf{x} \in \mathbb{R}^p$, the assumption of GMM is that $\mathbf{x}$ may come from K possible Gaussian distribution with corresponding probabilities. The probability density function of $\mathbf{x}$ can be expressed as:

$$f(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P_k f\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

where P_k represents the prior probability of the kth Gaussian Component, $f\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ is the conditional probability density function of the kth Gaussian

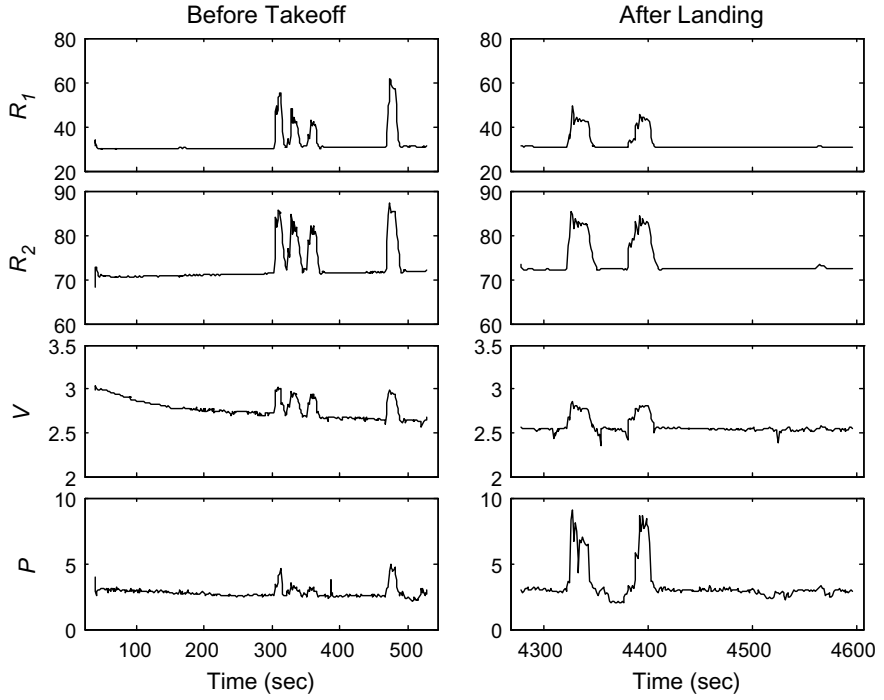


Fig. 4 The variables of a typical sortie before takeoff and after landing

Component with mean vector $\boldsymbol{\mu}_k$ and variance–covariance matrix $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\theta} = \{P_1, \ldots P_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$. Moreover, the probability density function $f\big(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\big)$ is given by:

$$f\big(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\big) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|} \exp\left[-\frac{1}{2}\big(\mathbf{x} - \boldsymbol{\mu}_k\big)^T \boldsymbol{\Sigma}_k^{-1}\big(\mathbf{x} - \boldsymbol{\mu}_k\big)\right]$$

and the prior probability P_k satisfies

$$P_k \geq 0, k = 1, \ldots, K \text{ and } \sum_{k=1}^{K} P_k = 1$$

Based on the Bayesian inference, the posterior probability of $\mathbf{x}$ belonging to the kth Gaussian component can be computed by

$$w_k = \mathbb{P}\big[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\mathbf{x}\big] = \frac{P_k f\big(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\big)}{\sum_{j=1}^{K} P_j f\big(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\big)}$$

Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ be a series of samples recorded every second. The parameters $\boldsymbol{\theta}$ can be estimated by the expectation and maximization (EM) algorithm which iterates the following E-step and M-step.

- E-step

$$w_{tk}^{(s)} = \frac{P_k^{(s)} f\left(\mathbf{x}_t | \boldsymbol{\mu}_k^{(s)}, \boldsymbol{\Sigma}_k^{(s)}\right)}{\sum_{j=1}^{K} P_j^{(s)} f\left(\mathbf{x}_t | \boldsymbol{\mu}_j^{(s)}, \boldsymbol{\Sigma}_j^{(s)}\right)}, t = 1, \ldots, T \text{ and } k = 1, \ldots, K$$

- M-step

$$P_k^{(s+1)} = \frac{\sum_{t=1}^{T} w_{tk}^{(s)}}{\sum_{j=1}^{K} \sum_{t=1}^{T} w_{tj}^{(s)}}$$

$$\boldsymbol{\mu}_k^{(s+1)} = \frac{\sum_{t=1}^{T} w_{tk}^{(s)} \mathbf{x}_t}{\sum_{t=1}^{T} w_{tk}^{(s)}}$$

$$\boldsymbol{\Sigma}_k^{(s+1)} = \frac{\sum_{t=1}^{T} w_{tk}^{(s)} \left(\mathbf{x}_t - \boldsymbol{\mu}_k^{(s+1)}\right)\left(\mathbf{x}_t - \boldsymbol{\mu}_k^{(s+1)}\right)'}{\sum_{t=1}^{T} w_{tk}^{(s)}}, \quad k = 1, \ldots, K$$

where $w_{tk}^{(s)}$ denotes the posterior probability that the sample recorded at the tth second comes from the kth Gaussian component at the sth iteration, and $P_k^{(s+1)}$, $\boldsymbol{\mu}_k^{(s+1)}$, and $\boldsymbol{\Sigma}_k^{(s+1)}$ are the prior probability, mean vector, and variance–covariance matrix of the kth Gaussian component at the $(s+1)$th iteration.

There are two issues for estimating GMM parameters by the EM algorithm: the need of selecting the number of components and the requirement of careful initialization. Figueiredo and Jain [8] proposed an algorithm which can automatically determine the number of components and is not sensitive to initialization. However, Yang et al. [9] demonstrated an example showing that this algorithm may stop with too many Gaussian components. As Fig. 4 displays that there are two Gaussian components, the stable state and transitions, we use the EM algorithm described above with two Gaussian components. Also, in order to avoid being influenced by initialization, we repeated the EM algorithm 100 times with random initial conditions and took the result that maximizes the likelihood function in (1).

It is noted that there may be some sorties having more than two Gaussian components. As the rotating speeds during takeoff and landing are quite stable, we can draw the corresponding s chart for each rotating speed in the stable state to assess whether two Gaussian components are enough for modeling purpose. Clearly, the estimated standard deviation would be larger if there are more than two Gaussian components. The s charts of the two rotating speeds are shown in Fig. 5. We check the out-of-control sorties sequentially and found that all out-of-control sorties have

three Gaussian components except two of them which have four Gaussian components. Figure 5 also shows that the variability of the rotating speeds before takeoff is larger than that after landing. This is could be due to some initial checking procedure that involves testing the range of engine rotation before takeoff and the fact that the engine has fully warmed up after landing. Analysis based on data collected after landing would reveal the true condition of the engine. The resulting signal can be plotted to reveal impending failure which may also exhibit deterioration after some damages were sustained.

Let k^* denote the Gaussian component of the stable state. The representative statistics for the four variables of the stable state can be computed by the equations in the M-step, where $\hat{w}_{tk}$ is from the E-step and $\mathbf{x}_t$ now is the vector of all four variables. For the ith sortie, let $\mathbf{y}_i = \widehat{\boldsymbol{\mu}}_{ik_i^*}$ and $\mathbf{S}_i = \widehat{\boldsymbol{\Sigma}}_{ik_i^*}$. For the following analysis, we first check the homogeneity of the variance–covariance matrices and then compare the mean vectors through Hotelling's T^2 statistics.

Note that the homogeneity of the variance–covariance matrices in this case means that the variability structure among the four variables within each sortie is identical. The often used quantity for measuring the overall variability is the determinant of the variance–covariance matrix $|\mathbf{S}|$, which is called the generalized variance. Alt [10] proposed an approach in developing a generalized variance chart by using the first two moments of $|\mathbf{S}|$:

$$\mathbb{E}[|\mathbf{S}|] \pm 3\sqrt{\mathrm{Var}[|\mathbf{S}|]}$$

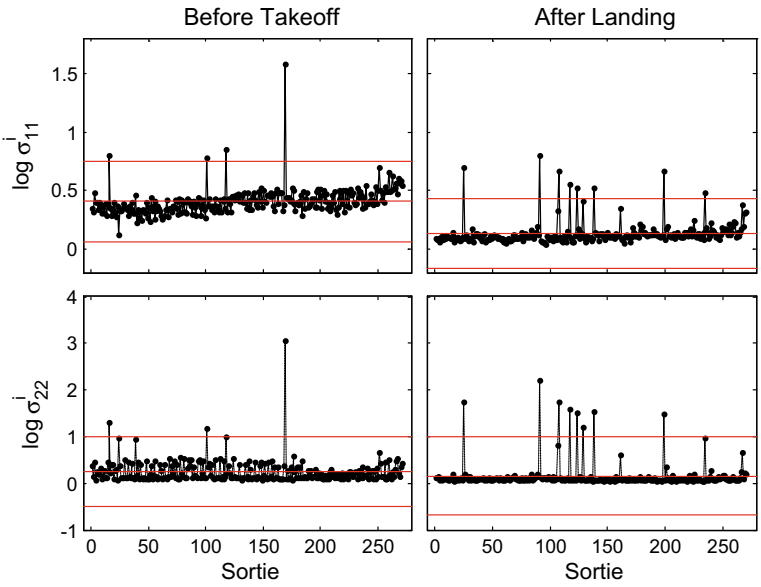


Fig. 5 The s charts of the two rotating speeds

However, this approach may not be appropriate because the distribution of the generalized variance is different from the Gaussian distribution. Goodman [11] proved that the distribution of the generalized variance is a constant times the product of p independent chi-square random variables with $2n, 2(n-1), \ldots, 2(n-p+1)$ degrees of freedom respectively, where n is the number of observations for calculating $\mathbf{S}$. As n is usually much greater than p, the logarithm of the generalized variance can be regarded as the sum of p independent and identical random variables, which is more suitable for the 3-sigma control limits. Since the logarithm of the generalized variance of each sortie can be regarded as an individual observation, we drew the individual control charts, where the variability is estimated by the moving range method, along with the histograms of the logarithm of the generalized variance in Fig. 6. It can be seen that the homogeneity assumption of the variance–covariance matrix is valid even though there are a few out-of-control points. Figure 6 also shows that the variability after landing is smaller than that before takeoff.

Next, we perform Phase I analysis for constructing a suitable control chart based on the Hotelling's T^2 statistics for monitoring the mean vector $\boldsymbol{y}$ of each sortie. Suppose that $m = 200$ sorties are used for Phase I analysis. Note that the between-sortie variability cannot be estimated by averaging the variance–covariance matrices, i.e. $\sum_{i=1}^{m} \mathbf{S}_i / m$ Since $\mathbf{S}_i$ represents the variability of the stable state within each sortie, it may be much smaller than the between-sortie variability. Instead, we consider using the mean square successive difference to estimate the between-sortie variability [12]. This approach is less sensitive to nonrandom patterns, such as trends and shifts, and thus it is useful for identifying nonrandom patterns in Phase I analysis.

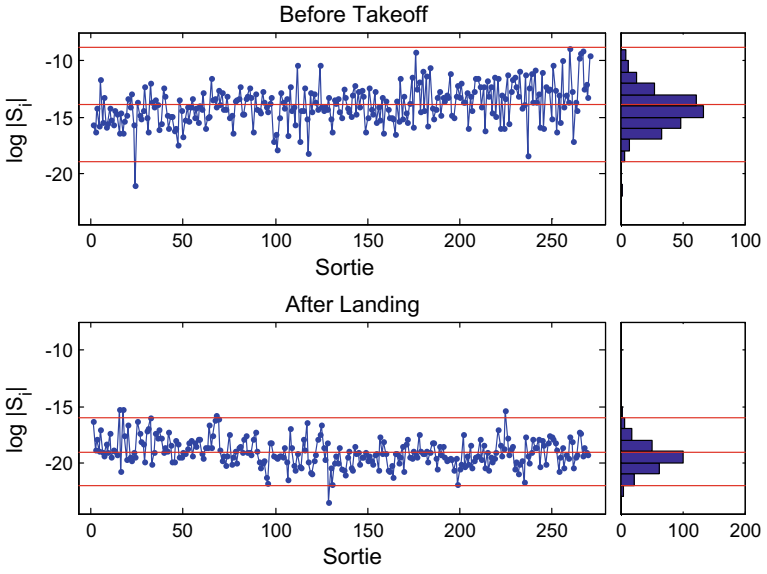


Fig. 6 The individual control charts of the logarithm of the generalized variance

Let $\mathbf{v}_i = \mathbf{y}_{i+1} - \mathbf{y}_i$ and $\mathbf{V} = \left[\mathbf{v}_1, \ldots, \mathbf{v}_{m-1}\right]^T$. The variance–covariance matrix can be estimated by

$$\mathbf{S} = \frac{\mathbf{V}^T\mathbf{V}}{2(m-1)}$$

The corresponding Hotelling's T^2 statistic is given by

$$T_i^2 = (\mathbf{y}_i - \overline{\mathbf{y}})^T \mathbf{S}^{-1} (\mathbf{y}_i - \overline{\mathbf{y}})$$

where $\overline{\mathbf{y}} = \sum_{i=1}^{m} \mathbf{y}_i / m$.

It is noted that the exact distribution of T_i^2 is hard to obtain and there are a few ways of determining the upper control limit (UCL). Sullivan and Woodall [13] proposed an approximate UCL

$$UCL = \frac{(m-1)^2}{m} \beta_{(p/2,(f-p-1)/2)}(1-\alpha)$$

where $f = 2(m-1)^2/(3m-4)$. Mason and Young [14, 26–27] suggested an adjustment to this approximation

$$UCL = \frac{(f-1)^2}{f} \beta_{(p/2,(f-p-1)/2)}(1-\alpha)$$

William et al. [15] observed that the asymptotic distribution of T_i^2 is a chi-square distribution with p degrees of freedom. Montgomery [16, 509] suggested using simulation to find the UCL since the cost is reasonable.

In this study, we found that the mean square successive difference approach would make the control charts overly sensitive which would give rise to more false alarms. As the condition of the engine was acceptable for most sorties in Phase I analysis, we use the pooled variance covariance matrix instead:

$$\mathbf{S} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T$$

The corresponding UCL is [17]

$$UCL = \frac{(m-1)^2}{m} \beta_{(p/2,(m-p-1)/2)}(1-\alpha)$$

Figure 7 displays the Hotelling's T^2 control charts for Phase I analysis with $\alpha = 0.0027$. It can be seen that there are a few individual out-of-control points and a shift from 78 to 95th sorties for both charts. For diagnosis purpose, the individual control charts for the four variables are plotted in Fig. 8. Clearly, the individual out-of-control

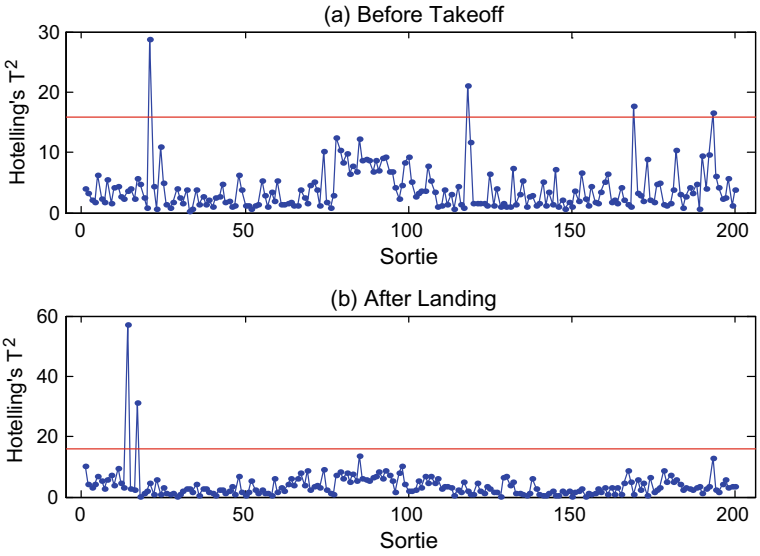


Fig. 7 The Hotelling's T^2 control charts for Phase I analysis

points are mainly due to unusual lubricant oil pressure, and the shift is because of relatively low rotating speeds. The control charts picked up some sorties that may contain potential fault. However, since these out-of-control signals appeared to be sporadic and there was no engine failure associated with these sorties used for Phase I analysis, the control charts constructed are acceptable.

The UCL for the Phase II analysis is given by

$$UCL = \frac{p(m-1)(m+1)}{m(m-p)} F_{(p,m-p)}(1-\alpha)$$

The Hotelling's T^2 control charts for the Phase II analysis with $\alpha = 0.0027$ are shown in Fig. 9. From Fig. 9a, the engine exhibited instability from the 225th sortie and experienced two extremely unusual sorties before the engine bearing failed. From Fig. 9b, there is a major shift starting from the 253rd sortie, and an extremely out-of-control point right before the bearing failed. The individual control charts for the four variables in Phase II are displayed in Fig. 10, where the center line and control limits are from the Phase I analysis. For the before takeoff part, the control charts indicate that the fluctuation of the Hotelling's T^2 control chart was caused by the two rotating speeds and the lubricant oil pressure, and the extreme signals before failure were because of the high lubricant oil pressure. On the other hand, for the after landing part, it can be seen that the shift resulted from the velocity of vibration, and the last extreme signal was due to the unusually high lubricant oil pressure. Moreover, by comparing individual control charts for both before takeoff and after landing parts, we found that the patterns of the two rotating speeds are similar, and

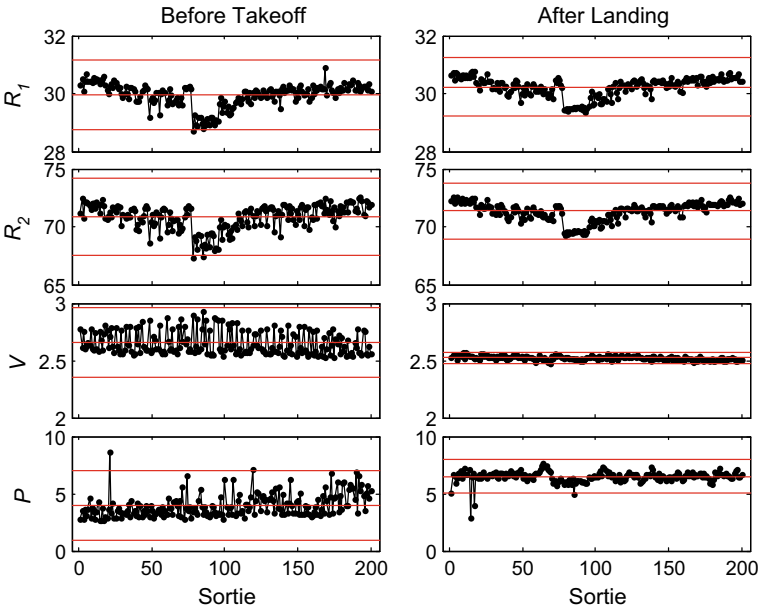


Fig. 8 The individual charts for Phase I analysis

the shift in the velocity of vibration was clearer after landing. As the engine bearing failed at the 271st sortie, the velocity of vibration and lubricant oil pressure could be the most related variable to the failure.

4 Conclusion

In this study, a series of graphical analysis of a set of in-flight data from an aircraft engine with the aim of detecting impending bearing failure and identify the key variables related to the failure have been presented. We first divided the sortie into three parts, and then applied the Gaussian mixture model to separate the stable state from transitions and extract the representative statistics. We proposed a control chart based on the logarithm of the generalized variance to check the goodness of fit of the Gaussian mixture model. From our example, this control chart is more suitable for the 3-sigma control limits. To monitor the between sortie variability with reasonable sensitivity, the Hotelling's T2 statistics based on variance–covariance matrix estimated by the mean square successive difference estimator and the pooled estimator were compared. In the current case, the latter is adopted as it has much fewer false alarms and capable to indicate the impending failure. Besides, we also found that control charts based on the after landing data would show clearer signals and follow-up actions can be taken. Finally, it should be noted that the four variables presented

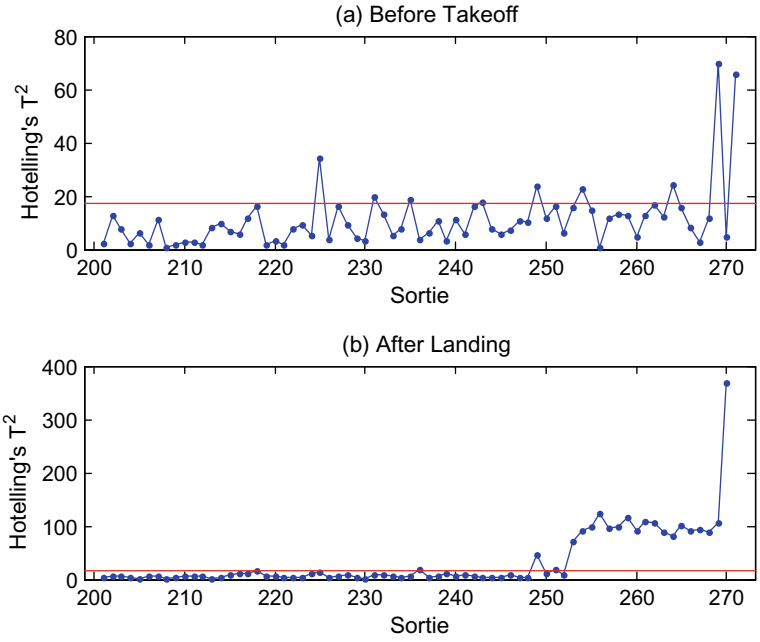


Fig. 9 The Hotelling's T2 control charts for Phase II analysis

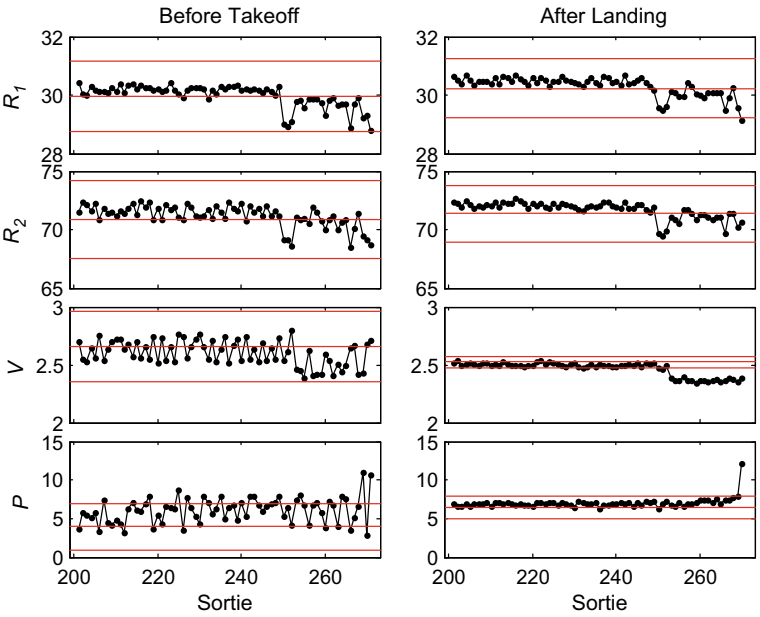


Fig. 10 The individual charts for Phase II analysis

here are the result of some pre-screening and engineering judgment. Through the analysis, the two variables, namely, the velocity of vibration and the lubricant oil pressure are further reaffirmed as the key variables associated to the failure.

References

1. Salam I, Tauqir A, Haq AU, Khan AQ (1998) An air crash due to fatigue failure of a ball bearing. Eng Fail Anal 5(4):261–269
2. Flouros M (2006) Correlations for heat generation and outer ring temperature of high speed and highly loaded ball bearings in an aero-engine. Aerosp Sci Technol 10(7):611–617
3. Orsagh RF, Sheldon J, Klenke CJ (2003) Prognostics/diagnostics for gas turbine engine bearings. In: Proceedings of IEEE aerospace conference, vol 36843, pp 159–167
4. Qiu H, Luo H, Eklund N (2009) On-board aircraft engine bearing prognostics: enveloping or FFT analysis? In: International design engineering technical conferences and computers and information in engineering conference, vol 48999, pp 1247–1252
5. Choi SW, Park JH, Lee IB (2004) Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. Comput Chem Eng 28(8):1377–1387
6. Yu J, Qin SJ (2008) Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. AIChE J 54(7):1811–1829
7. Yu J, Qin SJ (2009) Multiway Gaussian mixture model based multiphase batch process monitoring. Ind Eng Chem Res 48(18):8585–8594
8. Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell 24(3):381–396
9. Yang MS, Lai CY, Lin CY (2012) A robust EM clustering algorithm for Gaussian mixture models. Pattern Recogn 45(11):3950–3961
10. Alt FB (1984) Multivariate quality control. Encycl Stat Sci, Vol 6, Edited by Johnson and Kotz, Wiley, New York.
11. Goodman NR (1963) The distribution of the determinant of a complex Wishart distributed matrix. Ann Math Stat 34(1):178–180
12. Holmes DS, Mergen AE (1993) Improving the performance of the T2 control chart. Qual Eng 5(4):619–625
13. Sullivan JH, Woodall WH (1996) A comparison of multivariate control charts for individual observations. J Qual Technol 28(4):398–408
14. Mason RL, Young JC (2002) Multivariate statistical process control with industrial applications. Society for Industrial and Applied Mathematics
15. Williams JD, Woodall WH, Birch JB, Sullivan JH (2006) Distribution of Hotelling's T2 statistic based on the successive differences estimator. J Qual Technol 38(3):217–229
16. Montgomery DC (2020) Introduction to statistical quality control. John Wiley & Sons
17. Tracy ND, Young JC, Mason RL (1992) Multivariate control charts for individual observations. J Qual Technol 24(2):88–95

Novel Approach to Prognostics and Health Management to Combine Reliability and Process Optimisation

Dariusz Mazurkiewicz, Yi Ren, and Cheng Qian

Abstract Prognostics and Health Management (PHM) supports users with an integrated view of the health of any technical asset, and it consists of many different tasks based on data that are usually obtained from multisensory systems. The effective implementation of PHM does not, however, end with predicting remaining useful life (RUL). PHM has untapped potential to go beyond failure prediction and support of optimal maintenance actions and scheduling, along with logistics decisions. Both data captured by reliability systems and standard production data are generally used separately for different purposes. For higher effectiveness, these data have to be integrated in a combined approach. This can be achieved with the help of Digital Twin analytics that can support effective data use for parallel or combined purposes, such as classifying states, predicting failures or enhancing production efficiency. Furthermore, these seemingly independent concepts can be integrated into the same data collection approach. Previous studies have demonstrated that the afore-mentioned combined solution to classification and prediction challenges is yet only a standard approach to PHM, one that makes it possible to predict RUL, degradation track and optimal time to intervention. Consequently, a new solution is proposed, one that takes into consideration the possibility of intelligent and sustainable production in combination with online predictive maintenance and continuous process optimisation. The prediction of degradation and remaining useful life with the use of multisource data integration facilitates production process optimisation to gain additional use time. This, in turn, brings about incomparably greater financial effects than is the case with the traditional approach to PHM.

D. Mazurkiewicz (✉)
Lublin University of Technology, Nadbystrzycka 36, 20-618 Lublin, Poland
e-mail: d.mazurkiewicz@pollub.pl

Y. Ren · C. Qian
School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_23

1 Introduction

Industry 4.0 and related philosophies such as Maintenance 4.0 have radically changed the priorities in innovation and new technologies. Given the current trends in the development of industry and its economic transformation, with production companies playing the dominant role, the challenges and expectations for production engineering R&D are particularly enormous today. On the other hand, production engineering has all the necessary components and characteristics to play a significant role in innovative development, becoming both an inspiration for this development and its main driving force. Also, it can take into account sustainable development along with its social, economic and environmental challenges.

An analysis of recent trends in R&D demonstrates the growing significance of smartness. Technologies become smart; smart or even autonomous machines are implemented in smart production or assembly lines, and the maintenance of their elements is based on smart reliability. In addition, all these technologies and systems require smart sensors and also all together they create smart factories. Smart factories consist of smart machines and sensors or robotic platforms generating on the shop floor enormous amounts of data that are usually processed separately for different purposes, such as monitoring, reliability, logistics or production line management. As mentioned by Vališ and Mazurkiewicz [1], typical system operation data should be considered as a valuable source of information which can bring significant and tangible financial benefits at negligible costs if used properly in decision making. Unfortunately, a lack of understanding of the essence of data value or so-called information overload as well as difficulties with fast and effective analysis or processing of collected data due to the need of sophisticated tools, extensive knowledge and experience pose serious challenges to every company. No wonder that the engineering sector seeks to address these problems by innovation-supporting activities, where one of the information overload solutions for data mining is data dimensionality reduction. There also exist many other problems and challenges related to data collection, transfer, warehousing and analysis for effective decision-making, and they pertain to aspects such as data quality and the fact that data can be unlabelled, incomplete or uncertain.

Speaking only of reliability, it can be observed that production equipment must be maintained in constant operational efficiency, which can be ensured by effective monitoring and control of the technical condition of machines and devices [2–4]. It is particularly vital that potential failures of a given production system and its elements be known in advance and that the full operational efficiency of the system be restored in the shortest possible time, so as not to affect the manufacturing process. This requires the use of adequate maintenance strategies, techniques and tools, particularly in terms of data management, operations and executive actions. Several studies [5–13] have proved that the incorporation of advanced data analytics into reliability results in better decision-making support and higher operational efficiency.

One of the widely recognized and well-described practical approaches to reliability challenges is Prognostics and Health Management (PHM), which helps reduce

maintenance expenses effectively. PHM is of great importance in real-world applications as a tool for evidence-based planned maintenance actions and strategies. Machine health monitoring is considered a key component of modern manufacturing systems that have fully embraced the Big Data and Industry 4.0 revolution. According to [14], in contrast to top-down modelling provided by the traditional physics-based models, data-driven machine health monitoring systems offer a new paradigm of bottom-up solution for the detection of faults after the occurrence of certain failures (diagnosis) and the prediction of future working conditions and remaining useful life (prognosis).

Although prognostics has been applied in the field of maintenance for more than 10 years [15], the majority of its currently known real-world applications only address forecasting or remaining useful life (RUL) prediction. This is just a single facet of PHM among its many other highly valuable benefits. Data collected from different areas of the production system and its production floor for monitoring, logistics, PHM or process management are usually directed to separate databases and are also analyzed by different experts who work in isolation from one another. In particularly unfavourable circumstances, a considerable amount of data either remain unprocessed in information databases or are not processed as effectively as they need to be. This is a big mistake and a huge loss at the same time because it is well known that effective data processing and more advanced integration of data collected from different sources can improve the production process efficiency and operational management of the company. Importantly, too, predictions and decision models should dynamically react to changing production conditions and requirements of the Industrial Internet of Things (IIoT), which provides promising opportunities for developing novel industrial applications. The IIoT-based maintenance is an emerging research field with a great potential for industrial applications.

As noted by Koulali et al. [16] and Zhao et al. [14], the Industrial Internet of Things and data-driven techniques have revolutionized manufacturing by enabling computer networks to collect huge amounts of data from connected machines and to turn large machine data into actionable information. This information has to be transferred into knowledge, which will allow one to perform effective executive actions. In Industry 4.0, all these steps should be implemented in a smart system.

The above-mentioned problems, operational imperfections and limitations open up new opportunities and research challenges. First of all, the difficulties with the transition from data to knowledge and executive actions based on knowledge for different purposes require the development of new analytical tools and strategies with solutions for Big Data problems. Furthermore, the currently used solutions for failure predictions do not consider qualitative characteristics, sustainability or non-technical aspects of the production infrastructure. Therefore, they must be taken into account when developing smart analytical approaches. As an engineering discipline, PHM aims to provide users with an integrated view of the health of a machine or a general system, and its application consists of many different tasks, from sensing (usually with the use of multisensory systems) to prediction. Each task benefits from different modelling or data acquisition techniques; therefore, the typical PHM application does not necessarily depend on a single approach. For greater effectiveness, these

approaches have to be combined to go beyond reliability and thus take into account broader aspects of the problem. This can be achieved with the use of Digital Twin analytics to support effective data use for parallel or combined purposes.

To prove the validity of the above assumptions, a new solution is proposed, taking into consideration the possibility of intelligent and sustainable production in combination with online predictive maintenance and continuous process optimisation. For this purpose, first of all, an overview of the state of the art in PHM, its definitions, methods and R&D challenges is given (Chap. 2). Data-based approaches to PHM are also analysed (Chap. 3) in order to show the role of data acquisition and advanced data processing. Strengths and weaknesses of currently used data processing algorithms for PHM purposes are systematized. Since advanced and challenging data analytics needs cutting-edge solutions, the concept of Digital Twin is described (Chap. 4) as an effective PHM support tool for classifying states, predicting failures and enhancing production efficiency. After that, previous research results are presented (Chap. 5) to demonstrate that advanced and effective solutions to classification and prediction challenges—even when combined—are merely a standard approach to PHM. In light of the above, a new solution is proposed, taking into consideration the possibility of intelligent and sustainable production as a result of online predictive maintenance and continuous process optimisation integration. The prediction of degradation and remaining useful life with the use of multisource data integration and Digital Twin modelling facilitates production process optimisation to gain additional use time of a given asset. As a result, incomparably greater financial effects can easily be achieved in this way than with the standard approach to PHM.

2 Prognostics and Health Management

Modern machines and technical systems are extremely complex and comprise of many advanced interactive components and electronics. This complexity highlights the importance of technical system reliability. Failure of any technical element may result in a catastrophic system failure and thus generate huge costs. Therefore, the development of a usable system framework that is capable of early detection and isolation of an incipient fault of its components or subsystems is highly required. One of the best solutions in this respect is an effective PHM model that provides users with an easy-to-implement tool to monitor fault progression and support them in making decisions and creating maintenance strategies or schedules.

Widely used in the literature, the acronym PHM (Prognostics and Health Management) describes an important maintenance element of many systems and engineering products in which algorithms are used to detect anomalies, diagnose faults and predict remaining useful life. PHM is very often understood as an approach used for forecasting and reliability management or forecasting and management of the condition of a technical facility. As defined by Lee et al. [15], health management is the process of taking appropriate maintenance actions, on operational demand, and making accurate logistics decisions based on the results of available diagnostics and prognostic

resources. It focuses on assessing the impact of failures and minimizing negative results of any loss with maintenance management.

In the approach proposed by Li et al. [17], PHM is defined as a set of possibilities that cover both diagnostics and prognosis. Diagnostics refers to the process of detecting failures, whereas prognostics is the process of predicting the future condition of an asset or its RUL, depending on the current or historical conditions that are represented by e.g. measurement data. The diagnostics unit of PHM uses measurements for collecting necessary data, models and software to perform incipient fault detection. Models and software are also used by its prognostics unit for condition assessment and failure progression prediction. In this way, PHM provides users with the ability to determine the health state of any part, asset, subsystem or system.

PHM is also defined Sutharssan et al. [18] as a process of preventing failures and predicting reliability and remaining useful life. For engineering systems (especially technologically advanced ones) and strategic elements of production lines, this process is of vital importance because failure can not only cause serious damage to the system and environmental losses, but also pose a substantial threat to the safety of users or the environment. In economic terms, it can result in significant repair costs, unplanned operating costs in the future, or in crucial social and environmental losses. PHM is therefore widespread and at the same time recognized as an effective and practical approach to maintenance strategies and their challenges. Repair and maintenance costs can be reduced, for example by converting unplanned maintenance tasks into evidence-based scheduled preventive maintenance. An evidence-based scheduled maintenance strategy reduces inspection costs, the number of skilled workers required to complete the inspection, system downtime and lifecycle costs, as well as eliminates unscheduled maintenance actions and their impact on the environment and users.

As reported by Zhang et al. [19], PHM methods can be roughly classified as either physical model-based or data-driven. Similarly, one can also distinguish three types of RUL prediction models [20]: model-based methods, data-driven methods and hybrid methods. Data-driven methods attempt to acquire hidden knowledge from empirical data, infer about the current health state of an item of interest and predict its RUL. Data-driven methods can be classified as supervised and unsupervised, depending on whether the raw data are labelled or not. With the growing data overflow in industry causing Big Data problems and the widespread popularity of computing power as well as information and communication technologies, data-driven methods are considered as offering more opportunities in PHM applications. With the use of PHM data-driven maintenance strategies, reliability engineers are able to predict when equipment failure might happen and perform in advance maintenance actions to keep machines or systems in operation.

According to Rezaeianjouybari and Shang [21], PHM has emerged as a key technology to overcome the limitations of traditional reliability analysis and maintenance strategies. PHM focuses on utilizing sensory signals acquired from an engineered system to monitor health condition, detect anomalies, diagnose faults, and, what is more important, predict the RUL of a system over its lifetime. The health information

thereby obtained provides advance warning of potential failures, creating an opportunity to implement failure prevention measures by reducing system downtime and maintenance costs.

The condition of a technical object can be effectively assessed [22] as a degree of deviation or deterioration in relation to the expected typical operational performance that must be precisely defined to prevent potential failures. It is also necessary to determine which operating parameters contribute to that degree of deviation or degradation. There exist two different approaches to assessing the degradation or extent of deviation from the expected performance to assess system reliability and RUL forecasting via PHM: a data-driven approach (using statistical or machine learning methods) and a model-based approach.

PHM applications can be classified into two main categories [18, 20]:

- Real-time PHM, usually referred to as online PHM.
- And off-line PHM, which is deployed where system safety is not critical and the probability of failure is very low.

Most of the currently used machines or systems require real-time PHM (referred to as on-board health monitoring). Real-time PHM is also known as built-in self-test (BIST) or self-scanning lines [18]. PHM can be more effective if life cycle loads and other technological or reliability parameters are monitored in real time. This approach is particularly needed with critical machines and technological systems. PHM is also an effective tool for developing systems, machines or products to ensure their expected performance in compliance with relevant reliability requirements.

According to Lee et al. [15], PHM should be treated as an evolved form of the condition-based maintenance (CBM). CBM techniques are widely used to provide input for prognostic models in PHM and to support accurate and timely decision-making for preventing downtime. Given its ability to assess health status and predict the occurrence of failure and downtime, PHM is considered to be the foundation of any reliability system, when complemented with other techniques (self-maintenance, resilient and engineering immune systems).

In the context of PHM, a lack of representative high-quality data sets has obstructed the broad usage and adaptation of its approaches in industrial applications [23–26]. For effective PHM approaches, several sources of information are required, including historical, monitoring or modelling data. These data can be processed to obtain information or new knowledge with the use of a statistical or machine learning approach. In the statistical approach, predictions are made based on known or unknown underlying probabilistic distributions. As emphasised by Sutharssan et al. [18], under the parametric approach, parameters associated with probability distribution are calculated from data. Typically, data will represent healthy system performance under expected typical operating conditions. Healthy, or normal operating data can be defined by these parameters and are assumed to be a representative probability ability distribution, which is then used to detect anomalies and predict RUL. Once the system's healthy or normal operating data are defined by a probability distribution, new monitored data can be classified or applied for predictions using different methods based on the probability distribution.

For this approach historical data are necessary, which is not always feasible. In particular, the users of new machines or systems with no available historical reliability or usage data emphasize that the principal disadvantage of the data-driven approach is that under these circumstances predictions are either inaccurate or possible only after a longer period of data collection. In some instances, to overcome these limitations, one can use laboratory aging methods to obtain necessary data [27–29]. Nevertheless, the successful application of aging processes, accelerated load tests and semi-supervised strategies may require long time and expensive tests, and still fail to generate these data. Aging processes and accelerated load tests are applied to cause the investigated object to fail more quickly than under regular use conditions. The semi-supervised learning approach is used to define a healthy status of an object, assuming that new machines or technological lines would not fail for a certain period of time.

Despite numerous research achievements, there are still several challenges related to prognosis and health management applications, including:

- Anomaly detection, feature extraction, remaining useful life estimation, maintenance strategies development and implementation according to Industry 4.0 requirements [30].
- Research on heterogeneous unsupervised domain adaptation, particularly to complex physical systems, has been very limited, but has great potential to be impactful, particularly for advanced industrial applications [23–26].
- A lack of visualization tools for the dissemination of PHM information or decision-making support [15].
- The integration of PHM in the smart manufacturing paradigm should go beyond monitoring and data analysis for an individual component. This is challenging and requires a great deal of computational resources to determine and manage interactions between components, subsystems and systems. There is a lack of embedded sensors within manufacturing equipment that can be used to identify conditions showing changes in failure distribution. It is necessary to utilize the IIoT and predictive analytics to identify a state shift in the failure distribution, with a further development of the predictive maintenance policy that will be able to reschedule maintenance accordingly [31].

3 Data-Based Approaches to PHM

A typical monitoring or measurement system consists of sensors installed in a physical system in order to analyse, observe and control the production system or its assets in real time. There exist many types of professional instruments for acquiring and analysing signals from a machine or process, such as sensors, counters, or controllers that are designed to diagnose a given machine or its components. Having access to several data sources, including historical sets of healthy and failure data (i.e., labelled data), one can apply the supervised learning approach to predict an output for new

input data. Most PHM-related problems can be treated as supervised learning problems, and, as a result, techniques such as machine learning algorithms can be used to fuse the measurements for the same variable from multisensory, and many others may be successfully applied. Machine learning approaches, which can be classified as data driven, make predictions based on acquired data by converting the data into useful information that can be used in conjunction with current sensor data to provide future predictions. Depending on the type of algorithm, data-driven methods can be categorized into statistical data-driven methods and deep learning-based methods.

Data engineering is an important aspect of Maintenance 4.0. For various reliability and process data sources, adequate acquisition methods, feature and knowledge fusion strategies as well as architectures have been developed over the last few decades. As highlighted by Roemer et al. [32], the fusion of relevant sensor data, maintenance database information and outputs from various diagnostic and prognostic technologies has proven effective in reducing false alarm rates, increasing confidence levels in early fault detection and predicting time to failure or degraded condition requiring maintenance action. Generally, the objective of data fusion is to combine their respective information in the most diagnostically efficient method possible. Multisensory data fusion refers to intelligent processing of an array of several sensors that have cooperative, complementary and competitive qualities. It is estimated that there exist probably hundreds of techniques for performing data fusion, feature fusion or knowledge fusion [32].

As emphasized by Sankavaram et al. [33], depending on the type of information used, prognostic techniques can be categorized into three types: data-based time-to-failure, stressor-based and degradation-based. Time-to-failure data-based methods use failure time data to estimate the lifetime of a component (e.g. Weibull analysis). Stressor-based methods consider operating conditions such as temperature, humidity, vibrations, load, input current, and voltage. Degradation-based methods estimate and track degradation parameters and predict when the total degradation or damage exceeds a predefined functional failure threshold for the constructed monotonic health indicator (HI) (Fig. 1). These degradation parameters can be independent variables that are directly measured from the system or can be acquitted as a fusion of multiple parameters. Among several applications of this approach, the dynamic updating method of failure threshold depending on the confidence interval was proposed by Merh [34] and Li et al. [35] for zero RUL time estimation.

According to Yan et al. [20], in most existing RUL predictions, the estimated RUL was a determinate value and provided limited sensor-equipped machine status information, which was not conducive to enterprises making the corresponding maintenance decision. Therefore, a wider scope of application for advanced measurement techniques, together with smart maintenance supported by expert systems with its promising potential for PHM applications, is particularly well positioned to offer solutions to the following problems [24, 30, 36–38]:

- Automatic data processing of massive amounts of several multisensory system data.

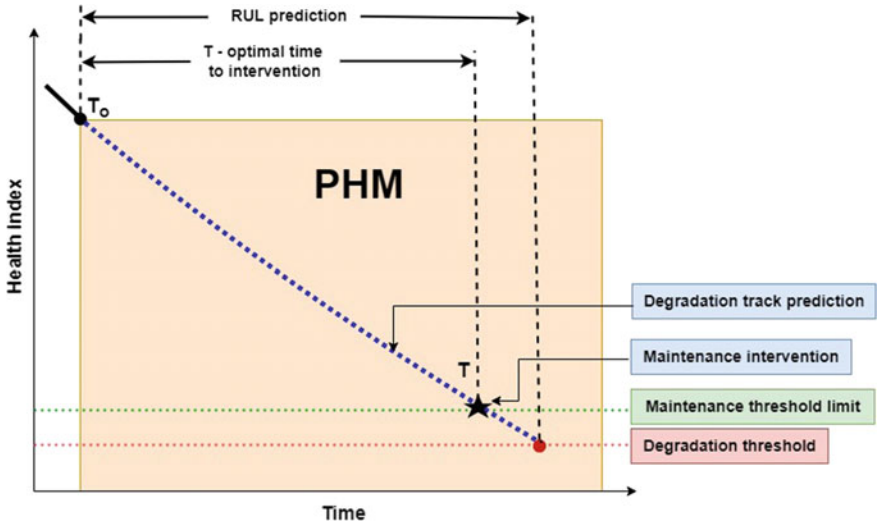


Fig. 1 PHM-based RUL prediction and optimal maintenance action scheduling

- Automatic extraction of useful features from high-dimensional, heterogeneous data sources.
- Ensuring good ability to distinguish functional and temporal relationships between and within the time series of condition monitoring signals.
- Knowledge transfer between different units.
- Precise RUL estimation to reduce unnecessary maintenance activities and eliminate potential risks, thereby improving sensor-equipped machine production efficiency and reliability.

There are a wide variety of algorithms in the field of PHM data engineering, depending on the application and available infrastructure. Each of them has several strengths and weaknesses when considered as a PHM solution [15, 20, 39–43], including:

- The Fourier transform, which usually presents signals with good spectrum resolution, is not suitable for nonstationary signals.
- Principal component analysis is able to reduce multidimensional data sets to lower dimensional data sets, but its performance varies for different applications.
- Logistic regression is appropriate when the output is considered between 0 and 1, but it is not appropriate when the output is unbounded.
- The Kalman filter is capable of both estimating the current state and predicting the future state, but it only works with linear systems and Gaussian noise.
- The use of decision trees allows for good visualization, easy interpretation and quick analysis ability for decision making, but their structure design and interpretation require experience and knowledge.

- Although the Support Vector Machine (SVM) is efficient for large datasets and real-time analysis, there is no standard method for choosing the kernel function, which is the key process for SVM. The use of SVM requires both experience and good knowledge.

It is well known that the main objective of maintenance engineers is to find a PHM architecture that can start inference and executive actions with very little prior knowledge about machine behaviour. This means that such architecture should be able to detect novel behaviours and learn from new incoming data. Also, it should be able to do the following [44]:

- Allow simultaneous monitoring of machines distributed worldwide.
- Help collect data in a structured way, offering labels associated with each observation, i.e., low-frequency data associated with high-frequency data.
- Provide an early indication of anomalous and novel behaviours.
- Learn from experience in different contexts, as well as perform different functions depending on the specific component and the whole machinery.

Most of the prognostic approaches reported in the literature consider one or two categories of data. More significantly, they are component-centric and primarily focus on predicting the remaining useful life of one particular component in isolation [33]. Although the approach to fault prognosis in coupled systems involving the combination of three types of data: failure time data, static environmental and status parameter data as well as dynamic data, was proposed as a novel framework for data-driven, semi-supervised and partially online PHM applications in industries [44], it does not include RUL estimation.

Despite their several disadvantages which outnumber the advantages, machine learning methods have attracted intense interest for PHM applications due to their representation power, automated feature learning capability and a very good performance in solving complex problems with the use of multisensory data. Li et al. [35] observe that given the increasing amount of data collected by sensors in industrial production, data-driven approaches hold great promise in leveraging monitoring data to enable the prediction of performance degradation in complex mechanical systems.

Measurement data collected with the appropriate sampling frequency are most often a time series of observations showing the change in the examined phenomenon over successive periods. There are two basic methods of time series analysis in maintenance [13, 45]. The first one is related to the study of relationships between elements of a given time series (correlation analysis). The aim of this approach is to estimate the correlation function using parametric methods. The second approach is related to the analysis of frequency characteristics of a series (spectral analysis of time series). Various spectral, asymptotic and functional techniques are used for this purpose. The time series is also characterized by four basic factors: trend, seasonality, cyclicity, and randomness. Nevertheless, the time series does not always contain all these four factors. Several studies assume the existence of random factors. Another important factor is the stochastic trend that occurs in the time series as a result of the integration

of past disturbances. The time series can also be stationary or nonstationary. Time-series data represent dependencies [19] that are usually crucial for fault diagnosis and may also contain cross correlations amongst multiple, multisensory measurements.

As mentioned by [17], a number of applications have been developed in recent research for asset-specific modelling and predictions with the use of time series. Consequently, there is some inconsistency in the understanding of key concepts for designing prognostic systems. To progress from the application-specific solutions to structured, consistent and efficient implementations of the PHM system, the development and use of a suitable methodology is essential. Such methodology should address the following high-level requirements [4]:

- It should be unambiguous, that is, the concepts and terminology used should be defined precisely, without being open to multiple competing interpretations.
- It should be comprehensive, that is, it should cover all essential steps in developing a PHM system.
- It should be pragmatic, that is, researchers and practitioners alike should be able to apply the methodology in a straightforward fashion.

In addition to that, according to Sankavaram et al. [33], the time series-based approaches to prognostic health management are component-centric and do not make use of widely available data in archived databases. Nevertheless, with the currently used tools, this limitation can be overcome [46]. Notwithstanding the above, many publications draw a conclusion that real-world tests and implementations are missing outside of academic studies. Han et al. [47] emphasize that modern manufacturing systems comprise of complex manufacturing equipment, processing personnel and workshop environments. Numerous studies have investigated the RUL of manufacturing equipment based on machine operating data in the industrial Internet environment. However, these studies mainly focus on the performance state degradation of a single machine in complex systems. No system-level RUL target has yet been established to dynamically characterize, analyse and optimize the overall operational state of a manufacturing system. The difficulty with measuring functional failures lies at the system level. The degradation process of mechanical equipment in a complex manufacturing system is not an independent stochastic degradation process; rather, it is affected by many factors. Not only is it difficult to establish a physical model representing the above-mentioned decisive factors for RUL estimation around a single device, but it is also hard to guarantee the accuracy of estimation results. Therefore, the RUL of a manufacturing system depends not only on the performance state of each manufacturing equipment but also on the ability of the whole manufacturing system to complete a given production task. Tao et al. [48] claim that under the new production paradigm known as smart manufacturing, a manufacturing factory must be flexible enough to allow multiple variations of production sequences as well as to adapt changes in the manufacturing system to new customized/individualized product offerings. According to Leng et al. [49], various advanced manufacturing strategies are employed for increasing manufacturing system automation to achieve smartness. Especially with customized products, traditional manufacturing systems face

a considerable challenge to realize customized production with large-scale production efficiency, because they cannot integrate information, equipment and services on time.

Under the Delphi-based scenarios for the 2030 maintenance in digitalised manufacturing [50] and according to several other future Industry 4.0 trends [51, 52], real-time control, predictability, efficiency and safety are recognised as key challenges of Maintenance 4.0. It has been proved that the effective implementation of PHM is one of the possible solutions. The PHM process does not end with remaining useful life estimation. It goes beyond failure predictions and, additionally, supports optimal maintenance actions and scheduling together with logistics decisions. In all of them, available resources are taken into account, together with the operating context and the sustainability consequences of different actions. That is why, Lee et al. [15] describe health management as a process of taking timely and optimal maintenance actions based on the outputs of diagnostics and prognostics, available resources and operational demand. PHM helps to assess and minimize the operational impact of failures, while at the same time taking into account the control of potential maintenance costs.

4 Digital Twin as PHM Support

Observations made by Van Horenbeek and Pintelon [53] show that even though condition-based maintenance takes advantage of the known state of components, setting a degradation threshold beyond which preventive maintenance is carried out is not always an optimal solution compared to predictive maintenance, especially when considering interdependent multicomponent systems. In addition to current degradation information, predictive maintenance also makes use of predictive information in the form of the RUL of components to optimally schedule maintenance actions, while condition-based maintenance only uses current component state information. Proactive maintenance decisions can be made based on predictive information, which results in a dynamic maintenance schedule. Moreover, predictive information makes it possible to include component interdependencies in the maintenance schedule. To do so, adequate time series modelling is necessary [54]. This, however, is not the only challenge in that regard. Another perspective to address Maintenance 4.0 and PHM challenges is to create a simulation environment and adapt the knowledge obtained therefrom to real-life applications. Such approach is possible with the use of Digital Twin (DT) concept. This research and development direction is particularly interesting, as data will most likely be sufficient in the source domain.

Digital Twin is capable of creating connections between physical and virtual spaces for smart maintenance solutions and is considered one of the strategic priorities shared by all major manufacturing concepts, such as Industry 4.0 and the Industrial Internet of Things or Machine to Machine Communication. Sensors and data transmission technologies are now increasingly used in production plants to collect data throughout different stages of a product lifecycle, including product design,

manufacturing, distribution, maintenance, and recycling. Digital Twin analytics can support effective data use to classify states, predict failures and enhance production efficiency. According to Tao et al. [48], Kritzinger et al. [55], and Barricelli et al. [56], despite the increasing popularity of DT research, no efforts have been made to review DT applications in industry. The integration of DTs and services is a promising research direction. Not only can DTs enable new services, but they can also enhance the existing services by supplying new data. Many research problems such as maintenance scheduling or service optimization should be addressed in the future paradigm of DT-driven service integration. As summarized by Qi and Tao [57] and Tao et al. [48], the DT-driven PHM is characterized by the following:

- It shows great advantages over traditional PHM methods in terms of four aspects, i.e., model, data, interaction, and decision making. The DT-driven PHM holistically merges physical data and virtual data, real-time data and historical data, as well as enables data fusion. In this way, it corresponds to the broad trend that smart manufacturing is driven by big data.
- It integrates four dimensions of modelling (i.e., geometry, physics, behaviour, and rule modelling) to depict a practical situation more accurately.
- It connects physical and virtual spaces. In this way, not only can the physical entity be better controlled, but the virtual model can also be progressively optimized and upgraded.
- It helps maintenance decision making to be driven by high-fidelity virtual models on top of traditional optimization algorithms, leading to a more rational maintenance strategy.

Whatever the implementation or execution method, data-based predictive maintenance has nowadays become a new trend in prognostics and health management (PHM) for complex equipment. Data collected from different sources and monitoring systems of different forms and contents require the use of appropriate processing methods. The more accurate the data are and the more effective the processing method is employed, the better the prediction results might be achieved. Unfortunately, there exist [54] a number of several identified limitations to this approach, including nontechnical or sustainable aspects which have to be taken into consideration when designing decision support models. Digital Twin analytics can support effective data use to classify states, predict failures and enhance production efficiency.

5 Predictive Maintenance of Machine Tool System Combining Process Optimization

Innovative manufacturing companies usually adopt maintenance systems within their flexible production structures. It can be noticed that every machine (be it designed, added, removed or updated) has its own degradation process and that rebuilding system-level maintenance scheduling according to a different system structure leads

to huge computational complexity. In addition, machines consist of multiple components that are predisposed to cascading failures, where one possibility is that the failure of a component will lead to more workloads of other components. This loading dependence can result in failure propagation, causing the whole production line or system to be more susceptible to failure and making the maintenance decision-making process more difficult. Thus, maintenance becomes even more important because it is likely to benefit from a systematic prognostic and health management (PHM) framework. Not only can the system change its structure for flexibility and responsiveness, but it can also update the health prognosis based on monitored degradation signals in real time. Therefore, the novel transformation of traditional maintenance systems into smart maintenance should also take into account the requirement to be more adaptive, which, additionally, creates a need for further advances in research and development. There are very few studies investigating flexible structures in terms of maintenance optimization. An adaptive multiunit maintenance policy for sequential different structures can provide an important basis for the generalization and application of adaptive manufacturing systems. Most existing maintenance solutions focus on a manufacturing system with a fixed structure, which means that each dedicated strategy or single action is individually designed for each element of the system structure. As a result, maintenance strategies for adaptive manufacturing systems are still missing, and one of the research challenges is to develop actionable models. The literature review shows that several studies have made appreciable progress with regard to model creation, diagnosis analytics, problem identification, and RUL prediction. Given that operations, process optimisation and maintenance become analytics-driven, it is also important to develop more efficient models that provide end-users with actionable outcomes as automatic executive actions. As demonstrated by Yucesan et al. [30], instead of merely detecting that the bearing in a machine is starting to fail or that the bearing has 3 more months of useful life, analytics should be able to support decisions regarding measures to be taken to extend RUL and identify the root cause of the rate of unwanted degradation.

According to the maintenance theory, a PHM system should accomplish various functions based on a specific single component, which means that the functions should be performed separately, depending on the analysis goal. In addition to that, most definitions or studies consider maintenance and logistics to be the only beneficiaries of the PHM process. Nonetheless, this perspective is changing, and some researchers are looking for solutions that would combine different functionalities. In Lindström et al. [58], the authors investigate the possibility of intelligent and sustainable production that would combine and integrate online predictive maintenance and continuous quality control. The rationale for combining and integrating the two aspects is that continuous quality control can provide input to online predictive maintenance whenever no there are signs of maintenance problems, inadequate output is produced and process parameters cannot be adjusted to meet output specifications. Lindström et al. [58] effectively demonstrate that continuous quality control can be used as an additional indicator for maintenance needs, which may also involve check-ups and verification of the monitoring sensors and their function. Furthermore,

these two concepts can be successfully integrated into the same data collection, monitoring and analytics platform, as they are both based on sensor data combined with additional data modelled to obtain an input for making decisions. For the combined predictive maintenance of different types of technical objects Lee et al. [59] proposes that AI-based algorithms are applied to monitor the cutting tool and the spindle motor, which are two critical machine product families.

The objective of the project titled "Innovative measurement technologies supported by digital data processing algorithms for improved processes and products" is to ensure predictive maintenance of machine tool system and process optimization, while at the same time combining different functionalities of data engineering. This approach was inspired by several recent studies on machine multi-sensor systems and signal processing to develop a model for supporting the decision-making process in determining the service life of a cutting tool [12, 13, 46, 60–62]. One of the developed models [13, 62] supports the identification of parameters affecting the cutter condition. The predictive ability of the obtained model was assessed with the use of classification quality assessment indicators. The proposed method proved to be a simple way of cutter state identification. The results were promising, which was confirmed by the prediction model quality indicators. The obtained sensitivity was 0.98 and the false alarm rate was 0.0193. The proposed analytical solution, which was verified using real data from an industrial machine tool, can be used as part of the system for determining the wear rate of a cutting tool in the production process by acoustic signal analysis. Each task produced very good modelling results, even though it benefited from different modelling or data acquisition techniques.

Another inspiring approach to affective transition from raw industrial data to knowledge-based executive actions without human action was proposed by Kozłowski et al. [10], where the remaining useful life prediction was made via the combined use of Support Vector Machine (SVM) as a classification tool and Auto Regressive and Integrated Moving Average (ARIMA)-based identification (Fig. 2).

The results demonstrated that the use of historical data for the development of the SVM classifier followed by identification and prediction based on current monitoring data led to obtaining the required information with fewer errors as well as at lower costs and in a shorter time than before. This effective solution to classification and prediction challenges—even though combined—is only a standard approach to PHM, one that makes it possible to predict RUL, degradation track and optimal time to intervention (Fig. 3).

By predicting the degradation and remaining useful life with the use of multisource data integration, we are able to optimize the process to gain additional use time (T_3, Fig. 4).

In this way, any necessary maintenance intervention may take place later, which provides some added value to the maintenance process. Obtained modelling results will be ready for implementation in an expert system constituting the basis of a Computerised Maintenance Management System (CMMS). Based on the PHM functions such as current health assessment and RUL prediction, self-maintenance or self-optimisation will be also achievable. The terms self-maintenance and self-optimisation refer to the ability of a smart machine to perform regular checks by

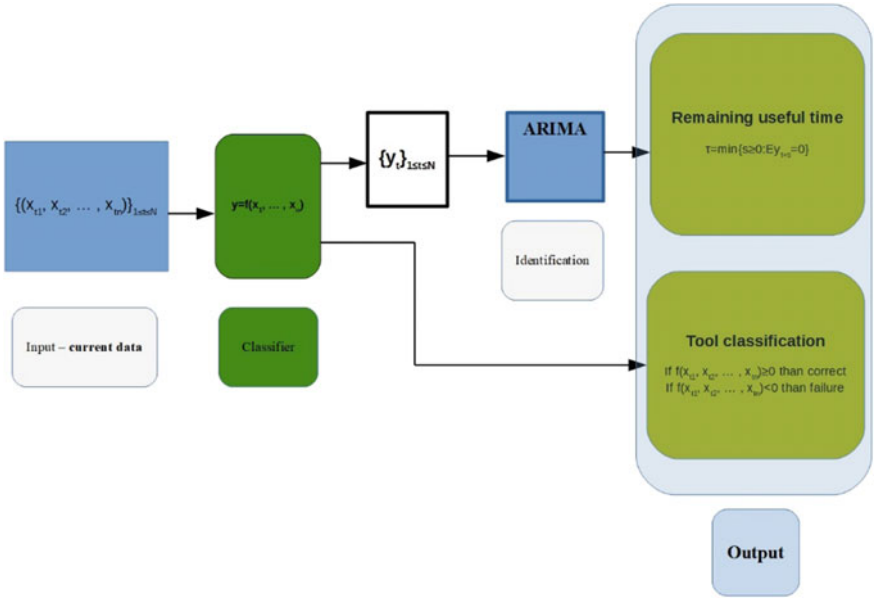


Fig. 2 RUL prediction and tool classification based on process data [10]

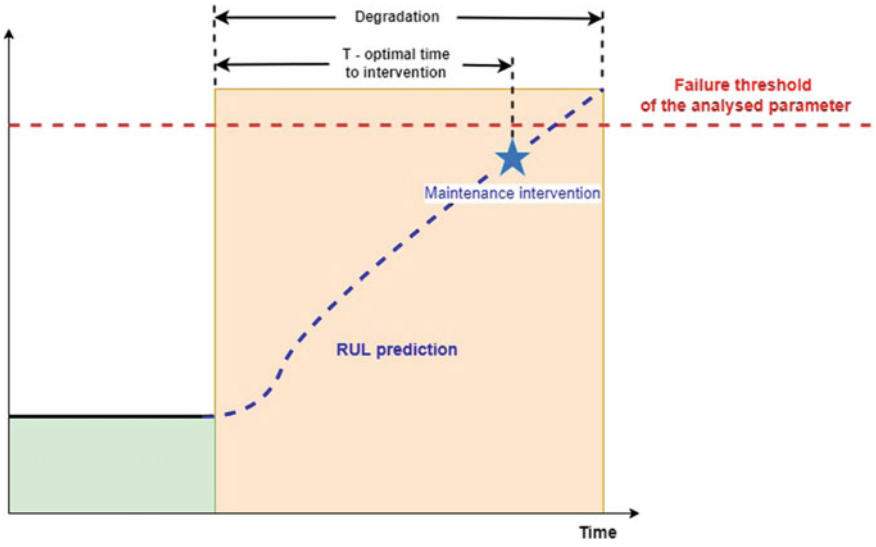


Fig. 3 Typical PHM-based RUL prediction and optimal maintenance action scheduling

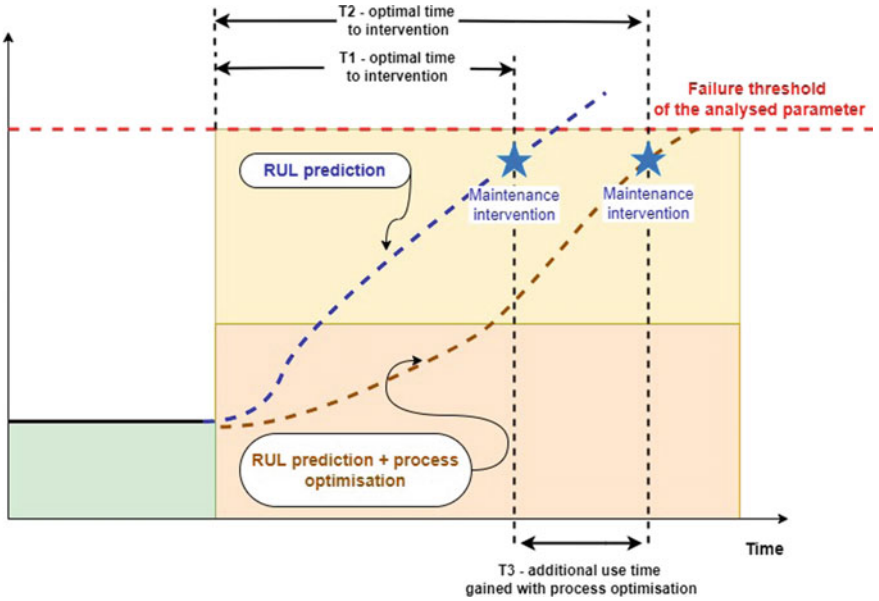


Fig. 4 PHM-based RUL prediction and optimal maintenance action scheduling with and without process optimisation

itself, detect possible anomalies and make immediate repairs if necessary by using stocked spare parts to avoid potential catastrophic loss. If stocked parts are unavailable, the self-optimisation part of the Digital Twin will be able to gain additional use time.

The above concept is under investigation within the framework of the PM/SP/0063/2021/1 project titled "Innovative measurement technologies supported by digital data processing algorithms for improved processes and products", launched in 2022 and financed by the Ministry of Education and Science (Poland) as part of the Polish Metrology Programme.

The aim of this project, which combines applied metrology and big data processing techniques with artificial intelligence and Digital Twin modelling (Fig. 5) as well as exploits the complementarity of resources and competences between consortium partners, is to develop innovative measurement techniques for industry in the field of metrology.

The primary goal of the research is to provide PHM and estimate RUL of components of technological machines and production systems. In addition to that, significant financial benefits are also expected, such as reduced operating and maintenance time and costs, as well as extended remaining useful life. Real-world data obtained from multisensory measurement systems will also help speed up the technology transfer to industrial applications, which is still very limited.

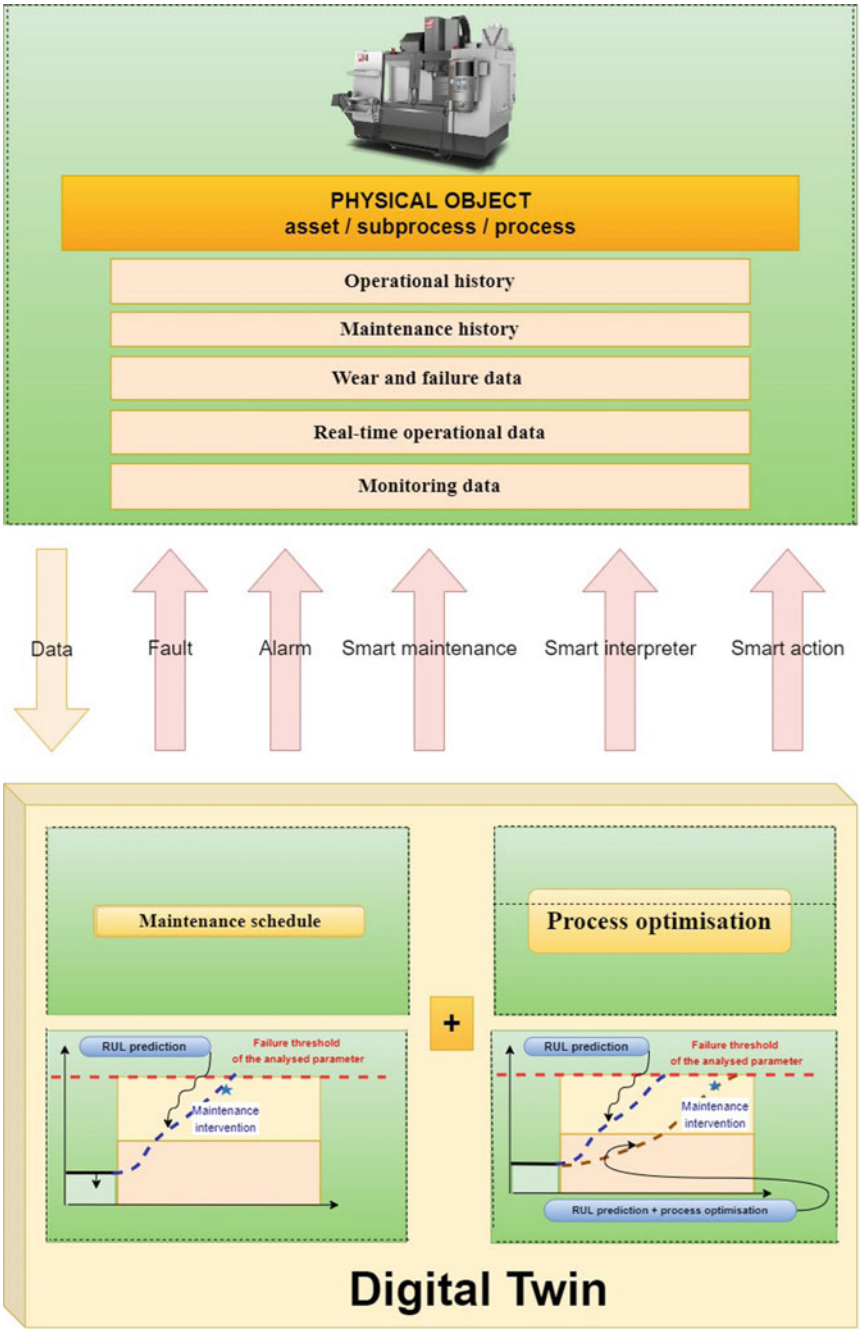


Fig. 5 Proposed digital twin concept for PHM with additional process optimisation to extend RUL of a process or its element

6 Conclusions

The state-of-the-art analysis demonstrates that recent advances in multi-sensor technologies and data-driven methodologies with the use of data engineering and Digital Twin simulations have significantly enriched the contents of prognostics and health management, including predictive maintenance. Still, more advanced analytic tools are necessary to adaptively and automatically extract information that is hidden in real-time measured streams, in order to perform executive actions combining different purposes. According to the concept described in this paper, PHM is considered as a holistic approach to the effective and efficient management of system health and process optimization. Generally, PHM applications require using more than one algorithm for different tasks, such as anomaly detection, parameter isolation, degradation parameter trending, damage estimation, and RUL prediction. Therefore, different types of algorithms can be employed to achieve these individual tasks. Under the proposed PHM concept, an added value can be achieved as a synergy effect of combining smart maintenance and advanced process management with the use of multisensory data and Digital Twin.

Several further studies are still needed to fill the identified development gaps. Some emerging future research and development challenges have been pinpointed by Lee et al. [15]. According to these authors, the concept of engineering immune system (EIS) will be the next generation of PHM, beyond self-maintenance and resilience systems. The EIS concept is explained by analogy with a biological immune system which protects against invasion and infection by identifying and killing pathogens. EIS can thus address machine maintenance issues in a highly complex and uncertain environment. The objective of EIS is to achieve efficient, near-zero breakdown performance with minimal human intervention. Technological visionaries believe that EIS should be robust in a diverse and dynamic environment, adaptive to learn and respond to new infections, able to retain memory to facilitate future responses, and autonomous in its self-controlled ability without the need for external control.

The demand for benchmarking application seems to be less urgent. The new architectures, algorithms, platforms, and frameworks described in this paper are now widely used for solving specific PHM problems that used to be unsolvable. Their variety creates a need for comparing workloads and final results.

Acknowledgements This work was supported by the individual research grant FD-20/IM-5/072/2022, of the scientific discipline Mechanical Engineering, Lublin University of Technology, Poland. This work was prepared within the project PM/SP/0063/2021/1 titled "Innovative measurement technologies supported by digital data processing algorithms for improved processes and products", financed by the Ministry of Education and Science (Poland) as a part of the Polish Metrology Programme. This work was partially prepared under the project "Digital Twin Based Fault Prediction and Maintenance Method for Intelligent Instruments"—Grant No. G2021177009L under the framework of the "High-end International Visiting Professor Project" through long distance collaboration.

References

1. Vališ D, Mazurkiewicz D (2018) Application of selected Levy processes for degradation modelling of long range mine belt using real-time data. Arch Civ Mech Eng 18:1430–1440
2. Zio E (2009) Reliability engineering: old problems and new challenges. Reliab Eng Syst Saf 94(2):125–141
3. Valis D, Mazurkiewicz D, Forbelska M (2017) Modelling of a transport belt degradation using state space model. In: 2017 IEEE international conference on industrial engineering and engineering management (IEEM). IEEE, pp 949–953
4. Li H, Wang W, Li Z, Dong L, Li Q (2020) A novel approach for predicting tool remaining useful life using limited data. Mech Syst Signal Process 143:106832
5. Levitt J (2009) The handbook of maintenance management. Industrial Press Inc.
6. Holgado M, Macchi M (2014) Exploring the role of E-maintenance for value creation in service provision. In: 2014 international conference on engineering, technology and innovation (ICE). IEEE, pp 1–10
7. Rokach L, Maimon O (2014) Data mining with decision trees. World Scientific
8. de Jonge B (2017) Maintenance optimization based on mathematical modeling. University of Groningen, Netherlands
9. Kosicka E, Kozłowski E, Mazurkiewicz D (2018) Intelligent systems of forecasting the failure of machinery park and supporting fulfilment of orders of spare parts. In: Intelligent systems in production engineering and maintenance–ISPEM 2017: proceedings of the first international conference on intelligent systems in production engineering and maintenance ISPEM 2017 1. Springer International Publishing, pp 54–63
10. Kozłowski E, Mazurkiewicz D, Żabiński T, Prucnal S, Sęp J (2020) Machining sensor data management for operation-level predictive model. Expert Syst Appl 159:113600
11. Tao F, Anwer N, Liu A, Wang L, Nee AY, Li L, Zhang M (2021) Digital twin towards smart manufacturing and industry 4.0. J Manuf Syst 58:1–2
12. Jasiulewicz-Kaczmarek M, Antosz K, Żywica P, Mazurkiewicz D, Sun B, Ren Y (2021) Framework of machine criticality assessment with criteria interactions. Eksploatacja i Niezawodność 23(2):207–220
13. Antosz K, Mazurkiewicz D, Kozłowski E, Sęp J, Żabiński T (2022) Machining process time series data analysis with a decision support tool. In: Innovations in mechanical engineering. Springer International Publishing, pp 14–27
14. Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX (2019) Deep learning and its applications to machine health monitoring. Mech Syst Signal Process 115:213–237
15. Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D (2014) Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. Mech Syst Signal Process 42(1–2):314–334
16. Koulali MA, Koulali S, Tembine H, Kobbane A (2018) Industrial internet of things-based prognostic health management: a mean-field stochastic game approach. IEEE Access 6:54388–54395
17. Li R, Verhagen WJ, Curran R (2020) A systematic methodology for prognostic and health management system architecture definition. Reliab Eng Syst Saf 193:106598
18. Sutharssan T, Stoyanov S, Bailey C, Yin C (2015) Prognostic and health management for engineering systems: a review of the data-driven approach and algorithms. J Eng 2015(7):215–222
19. Zhang L, Lin J, Liu B, Zhang Z, Yan X, Wei M (2019) A review on deep learning applications in prognostics and health management. IEEE Access 7:162415–162438
20. Yan J, He Z, He S (2022) A deep learning framework for sensor-equipped machine health indicator construction and remaining useful life prediction. Comput Ind Eng 172:108559
21. Rezaeianjouybari B, Shang Y (2020) Deep learning for prognostics and health management: state of the art, challenges, and opportunities. Measurement 163:107929
22. Sotiris VA, Peter WT, Pecht MG (2010) Anomaly detection through a Bayesian support vector machine. IEEE Trans Reliab 59(2):277–286

23. Che C, Wang H, Fu Q, Ni X (2019) Combining multiple deep learning algorithms for prognostic and health management of aircraft. Aerosp Sci Technol 94:105423
24. Fink O, Wang Q, Svensen M, Dersin P, Lee WJ, Ducoffe M (2020) Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng Appl Artif Intell 92:103678
25. Jlassi N, Cédrick B, Khlaief O, Medjaher K (2022) An Approach CPS for the smart monitoring of industrial systems. Omar and Medjaher, Kamal, An Approach CPS for the Smart Monitoring of Industrial Systems (July 26, 2022)
26. Tang L, Yang X, Gao J, Huang J, Cui J (2022) Adaptive Gaussian process regression based remaining useful life prediction of PEMFC incorporating an improved health indicator. In: 2022 IEEE 11th data driven control and learning systems conference (DDCLS). IEEE, pp 1080–1085
27. Gavrilov LA, Gavrilova NS (2005) Reliability theory of aging and longevity. In: Handbook of the biology of aging, pp 3–42
28. Mazurkiewicz D (2008) Analysis of the ageing impact on the strength of the adhesive sealed joints of conveyor belts. J Mater Process Technol 208(1–3):477–485
29. Celaya JR, Patil N, Saha S, Wysocki P, Goebel K (2009) Towards accelerated aging methodologies and health management of power MOSFETs (technical brief). In: Annual conference of the PHM Society, vol 1(1)
30. Yucesan YA, Dourado A, Viana FA (2021) A survey of modeling for prognosis and health management of industrial equipment. Adv Eng Inform 50:101404
31. March ST, Scudder GD (2019) Predictive maintenance: strategic use of IT in manufacturing organizations. Inf Syst Front 21:327–341
32. Roemer MJ, Kacprzynski GJ, Orsagh RF (2001) Assessment of data and knowledge fusion strategies for prognostics and health management. In: 2001 IEEE aerospace conference proceedings (Cat. No. 01TH8542), vol 6. IEEE, pp 2979–2988
33. Sankavaram C, Kodali A, Pattipati K, Wang B, Azam MS, Singh S (2011) A prognostic framework for health management of coupled systems. In: 2011 IEEE conference on prognostics and health management. IEEE, pp 1–10
34. Merh N (2019) Applying predictive analytics in a continuous process industry. In: Advances in analytics and applications, pp 105–115
35. Li Q, Ma Z, Li H, Liu X, Guan X, Tian P (2023) Remaining useful life prediction of mechanical system based on performance evaluation and geometric fractional Lévy stable motion with adaptive nonlinear drift. Mech Syst Signal Process 184:109679
36. Tiddens WW, Braaksma AJJ, Tinga T (2018) Selecting suitable candidates for predictive maintenance. Int J Prognostics Health Manage 9(1)
37. Ye Z, Si S, Yang H, Cai Z, Zhou F (2021) Machine and feedstock interdependence modeling for manufacturing networks performance analysis. IEEE Trans Industr Inf 18(8):5067–5076
38. Zhang K, Xia T, Wang D, Chen G, Pan E, Xi L (2023) Privacy-preserving and sensor-fused framework for prognostic and health management in leased manufacturing system. Mech Syst Signal Process 184:109666
39. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874
40. Jung W, Ismail A (2012) Prognostic and health management trend in automotive industry: an overview. In: Proceedings of conference Korea Society of industrial and systems engineering, pp 35–40
41. Wang J, Ma Y, Zhang L, Gao RX, Wu D (2018) Deep learning for smart manufacturing: methods and applications. J Manuf Syst 48:144–156
42. Tao F, Zhang H, Liu A, Nee AY (2018) Digital twin in industry: state-of-the-art. IEEE Trans Industr Inf 15(4):2405–2415
43. Nor AKM, Pedapati SR, Muhammad M, Leiva V (2021) Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses. Sensors 21(23):8020
44. Calabrese F, Regattieri A, Bortolini M, Gamberi M, Pilati F (2021) Predictive maintenance: a novel framework for a data-driven, semi-supervised, and partially online prognostic health management application in industries. Appl Sci 11(8):3380

45. Kozłowski E (2015) Analiza i identyfikacja szeregów czasowych. Wydawnictwo Politechniki Lubelskiej, Lublin, Poland
46. Kozłowski E, Mazurkiewicz D, Sęp J, Żabiński T (2022) The use of principal component analysis and logistic regression for cutter state identification. In: Innovations in industrial engineering. Springer International Publishing, pp 396–405
47. Han X, He Y, Wang Z, Cai Y, Dai W (2022) Remaining useful life prediction of manufacturing system based on fuzzy quality state task network. J Manuf Syst 65:233–243
48. Tao F, Tang Y, Zou X, Qi Q (2019) A field programmable gate array implemented fibre channel switch for big data communication towards smart manufacturing. Robot Comput Integr Manuf 57:166–181
49. Leng J, Liu Q, Ye S, Jing J, Wang Y, Zhang C, Ding Z, Chen X (2020) Digital twin-driven rapid reconfiguration of the automated manufacturing system via an open architecture model. Robot Comput Integr Manuf 63:101895
50. Bokrantz J, Skoogh A, Berlin C, Stahre J (2017) Maintenance in digitalised manufacturing: Delphi-based scenarios for 2030. Int J Prod Econ 191:154–169
51. Xu LD, Xu EL, Li L (2018) Industry 4.0: state of the art and future trends. Int J Prod Res 56(8):2941–2962
52. Bokrantz J, Skoogh A (2023) Adoption patterns and performance implications of smart maintenance. Int J Prod Econ 256:108746
53. Van Horenbeek A, Pintelon L (2013) A dynamic predictive maintenance policy for complex multi-component systems. Reliab Eng Syst Saf 120:39–50
54. Mazurkiewicz D (2021) Smart maintenance with time series modelling and digital twin
55. Kritzinger W, Karner M, Traar G, Henjes J, Sihn W (2018) Digital twin in manufacturing: a categorical literature review and classification. IFAC-PapersOnline 51(11):1016–1022
56. Barricelli BR, Casiraghi E, Fogli D (2019) A survey on digital twin: definitions, characteristics, applications, and design implications. IEEE Access 7:167653–167671
57. Qi Q, Tao F (2018) Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. IEEE Access 6:3585–3593
58. Lindström J, Larsson H, Jonsson M, Lejon E (2017) Towards intelligent and sustainable production: combining and integrating online predictive maintenance and continuous quality control. Proc CIRP 63:443–448
59. Lee WJ, Wu H, Yun H, Kim H, Jun MB, Sutherland JW (2019) Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data. Proc CIRP 80:506–511
60. Sun B, Pan J, Wang Z, Ren Y, Mazurkiewicz D, Jasiulewicz-Kaczmarek M, Antosz K (2021) Fatigue reliability analysis method of reactor structure considering cumulative effect of irradiation. Materials 14(4):801
61. Sun B, Yang X, Ren Y, Wang Z, Antosz K, Loska A, Jasiulewicz-Kaczmarek M (2021) Failure-based sealing reliability analysis considering dynamic interval and hybrid uncertainties. Eksploatacja i Niezawodność 23(2):278–284
62. Jasiulewicz-Kaczmarek M, Antosz K, Wyczółkowski R, Mazurkiewicz D, Sun B, Qian C, Ren Y (2021) Application of MICMAC, Fuzzy AHP, and Fuzzy TOPSIS for evaluation of the maintenance factors affecting sustainable manufacturing. Energies 14(5):1436

Review Paper

Current Status and Prospects of Reliability Systems Engineering in China

Yi Ren, Qiang Feng, Cheng Qian, Dezhen Yang, and Zili Wang

Abstract This chapter provides a systematic overview of the introduction and evolution of reliability systems engineering (RSE) in China, and the latest RSE development, including model-based RSE (MBRSE) and Reliability Digital Twin (RDT), are emphatically introduced. The chapter summarizes the establishment of the system architecture and conceptual models of MBRSE, fundamental theory and methodology of MBRSE with a V-model as the core of this approach, development of the MBRSE platform and RDT and the effectiveness of their implementations. Finally, the prospective trends in the development of RSE in China are outlined.

Keywords Reliability · Effectiveness · Model-based systems engineering · Reliability system engineering · Reliability digital twins

1 Introduction

Modern reliability engineering originated from the Unite States in the 1950s, and it has achieved remarkable effects after more than half a century of development. From the methodological perspective, it has undergone substantial changes from "passive improvement after the fact" to "active test to expose" and then to "active design to prevent". From the professional perspective, it has been developed from individual characteristics (such as reliability and safety) to comprehensive characteristics (such as the product's combat readiness, mission success, and effectiveness). From the technical perspective, it has been developed from statistics-based mathematical methods to Physics-of-Failure (PoF) based physicochemical methods, and then to data-driven and model-driven integration methods. From the work object perspective, it has been developed from hardware to a combination of hardware and software, from macroscopic scale to a combination of macroscopic scale and microscopic scale, from unit to system, and then it has been developed further to intelligent, networked and systematic advanced system.

Y. Ren · Q. Feng · C. Qian (✉) · D. Yang · Z. Wang
School of Reliability and Systems Engineering, Beihang University, Beijing, China
e-mail: cqian@buaa.edu.cn

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3_24

However, up to date, reliability has still been a worldwide challenge for complex equipment in modern engineering [1–4]. For instance, the underlying causes of many accidents, such as the derailment accident of the ICE-1 high-speed train in Europe in 1998 [5], the crash of space shuttle Columbia in 2003 [6], and a series of accidents of the Boeing 737 MAX from 2018 to 2019 [7], are related to the inadequate design and misestimation of reliability.

Different solutions have been proposed globally since the 1960s, especially in the US, Europe, and Japan, to tackle reliability problems existing in complex products, and design and manufacturing of high-quality products. For instance, in the US, the engineering specialty integration and concurrent engineering [8, 9] are proposed to provide engineering specialties related to reliability into the design process, so that the reliability and maintainability can be regarded as the design characteristics of products. The design, production, and support processes of products are conducted in parallel and interactively, thereby greatly improving the performance and quality of products and reducing the lifecycle cost. In Europe, ISO 9000–4 (dependability program management) is used to integrate the engineering and management specialties related to the inherent reliability of products for controlling the reliability over their lifecycles [10]. In Japan, the total quality management and robust design are proposed to consider the quality of products as the core and establish a scientific and efficient quality system [11]. Based on the aforementioned advanced technology and previous experience, Professor Weimin Yang, the pioneer and leader in reliability engineering in China, proposed the concept and theory of reliability systems engineering (RSE) with Chinese characteristics [12]. Compared with the relevant technologies developed abroad, China's RSE is an independent discipline system with a unified goal and quantifiable indexes that aims to deal with failures focusing on a common product. After approximately 30 years of development, the standards, procedures, and technologies (related to reliability, maintainability, and safety) of RSE have been gradually developed consistent with China's national conditions, and some remarkable application achievements have been made across Chinese industries.

This chapter is accomplished mainly on the basis of the authors' understanding of the theory of RSE and more than 30 years of practical experience from more than 10 military and civil fields, such as aerospace, shipbuilding, and industrial manufacturing, in China. Its motivation and contribution mainly include the following: (1) systematically reviewing the introduction and development of RSE and (2) introducing the latest development of RSE, namely, model-based RSE (MBRSE), and its future perspectives.

The rest of this chapter is organized as follows. Section 2 discusses the development and technological framework of RSE. Section 3 presents the conceptual and operational models of MBRSE. Section 4 explains the crucial technologies for MBRSE operation. Section 5 presents a platform and its application in MBRSE. Section 6 provides the conclusions and some representative directions.

2 Development of RSE

2.1 History of RSE

In China, engineers and scientists are committed to developing RSE and solving practical reliability problems for the past 27 years. The development trajectory of RSE in China is shown in Fig. 1.

Because China's industrial foundation, management model, and design culture are very different from those of western countries, it was not feasible to completely utilize foreign experience to develop reliability engineering technologies in China. The major challenges include the following: (1) There is a poor understanding of the scientific nature of quality engineering and the concepts of quality forming, leading to the fact that reliability does not lead to a prominent status. (2) Each characteristic in GQCs is introduced and independently developed with the lack of a main line, leading to a severe inconsistency problem. (3) Engineering methods and tools are limited to cause nonstandard work with a low efficiency and poor effect. (4) Advanced technologies such as prognostics and health management (PHM) started late. (5) The effectiveness of technology application is poor; the discrepancy between research and development (R&D) and application is large; the technical methods are less targeted and adaptable. Under this background, Professor Weimin Yang first developed the overall concept and fundamental theoretical framework of RSE by considering effectiveness as the goal and product failures as the core elements. The preliminary definition of RSE is as follows [12]: RSE is an engineering technique used to study the full lifecycle of a product and its actions in terms of failure mitigation. Apart from the dialectical relationship between the entirety of a product and its surrounding environment, RSE investigates the intercorrelation between the reliability and lifetime of a product and the surrounding environment, the failure occurrence and evolution, the laws to prevent, detect, mitigate, and eliminate

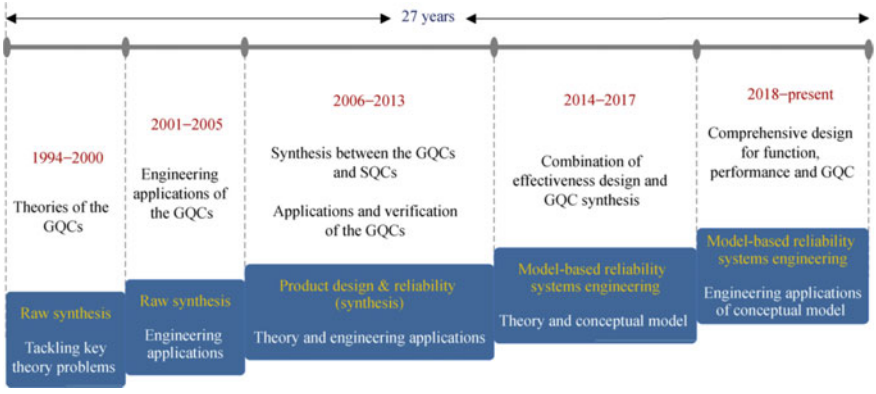


Fig. 1 Development trajectory of RSE in China

these failures, and a series of techniques and management activities to improve the reliability, prolong life, and enhance the effectiveness on the basis of various approaches, such as experimental research, field investigation, failure analysis, and maintenance. Professor Weimin Yang introduced a theoretical framework of RSE by analogy with the theories in medical engineering. The RSE of products was found to be extremely similar to the medical engineering of humans in many aspects, such as the "prevention and treatment of disease" and "good birth and healthcare conditions" [13]. RSE proposes a unified goal for integrating multiple specialties in reliability design. As such, RSE involves the application of systems engineering theory in reliability and the integration of reliability into the systems engineering process.

In 2005, the comprehensive quality view on three dimensions (CQVTD) pertaining to the overall characteristics, full lifecycle, and total system and the viewpoint of reforming quality engineering promoted by technology from manufacturing to designing the full lifecycle were proposed by Prof. Zili Wang, one of the authors of this chapter, to guide the development of RSE and to further strengthen the management and design [14]. In the CQVTD, the quality characteristics of a given product are divided into special quality characteristics (SQCs) corresponding to its function and performance and general quality characteristics (GQCs). At present, studies of GQCs in China mainly include reliability, safety, maintainability, testability, supportability, and environmental adaptability [13, 15]. Therefore, GQCs are sometimes referred to as the "six characteristics" in China. The CQVTD systematically explains the relationship between RSE and modern quality engineering and clarifies that RSE, with effectiveness as its goal and the synthesis of the "six characteristics" as its focus, aims to design comprehensive quality characteristics [15]. In 2005, with a focus on the prevention, diagnosis, and treatment of failures, the technical framework of RSE was further constructed in terms of its fundamental theory, fundamental technology, and application technology [14].

Based on the development of RSE, the authors of this study proposed a new definition of RSE in 2007: RSE is a synthetic cross-technology and management activity based on systems engineering theory that uses failures as its core elements and effectiveness as its goal and is designed to evaluate the laws of occurrence and the evolution of failures, including the stages of prevention, diagnosis, and repair, throughout the full lifecycle of a complex system. The definition of RSE was formally indexed in the Chinese Military Encyclopedia and General Introduction to Military Technology [16] (Shi 2007), and it suggests that RSE has been officially recognized as a discipline by the domestic engineering community in China. Since it was proposed, the development of RSE has been focused on solving the imbalance and inconsistency problems between the designs of GQCs and SQCs [15]. Synthesis within the GQCs, between the GQCs and SQCs, and between the technology and management of comprehensive quality characteristics must be continuously promoted to solve these problems.

In 2015, the core of RSE was further clarified as a combination of effectiveness design and GQC synthesis during the first international RSE conference. With the development of model-based systems engineering (MBSE), MBRSE was proposed in 2016 [17]. MBRSE integrates a large amount of work that is relevant to GQCs

to determine the failure laws on the basis of model evolution [18–21] and applies these laws to achieve the closed-loop mitigation and control (M&C) of failures by using models, such as the product, failure and environment models [22–24], as the core elements [17]. Such a process can be integrated into the MBRSE process of products.

2.2 Technological Framework of RSE

In the 1990s, customer requirements of a product evolved from the product's function and performance to effectiveness and cost-effectiveness ratio. Under these circumstances, effectiveness has become the synthesis of user concerns including product availability, dependability, and capacity. GQCs are an important basis for product effectiveness and a key factor in influencing lifecycle costs. By systematically considering the relationships between effectiveness and GQCs, three principles of RSE are identified: (1) global view, (2) systematic process, and (3) synthetic method. These principles make the RSE studies in China relatively different from the corresponding studies in foreign countries. For the "global view," MBRSE coordinates the functional/performance model groups and GQC model groups in terms of the product, function, and usage at the global level by adopting effectiveness as the goal. For a "systematic process," MBRSE is used to plan the model-driven reliability work throughout the full lifecycle of products on the basis of multidimensional failure logics such as failure prevention before product delivery and failure prognosis and diagnosis during operation. In the "synthetic method," MBRSE achieves data integration, process integration, and characteristic synthesis among the functional/performance model groups and GQC model groups to further achieve GQC technology and management synthesis, driven by failure identification and mitigation.

On the basis of the above-mentioned principles, the current technological framework of RSE [14] involves three levels: fundamental theory, basic technology, and applied technology, as shown in Fig. 2.

1. Fundamental theory

RSE was developed on the basis of failure recognition theory that elucidates failure mechanisms and determines failure laws to support failure prevention, control, and maintenance technologies [14]. Failure recognition theory integrates the physics of failure (PoF) that occurs under the load response and physicochemical process [25], the logic of failure (i.e., statics, dynamics, and emergence logics) [26, 27], and human error, which is categorized under performance effect and ability limitation [28]. Its mathematical and physical fundamentals are highly related to the certainty and uncertainty theories and their combination [23].

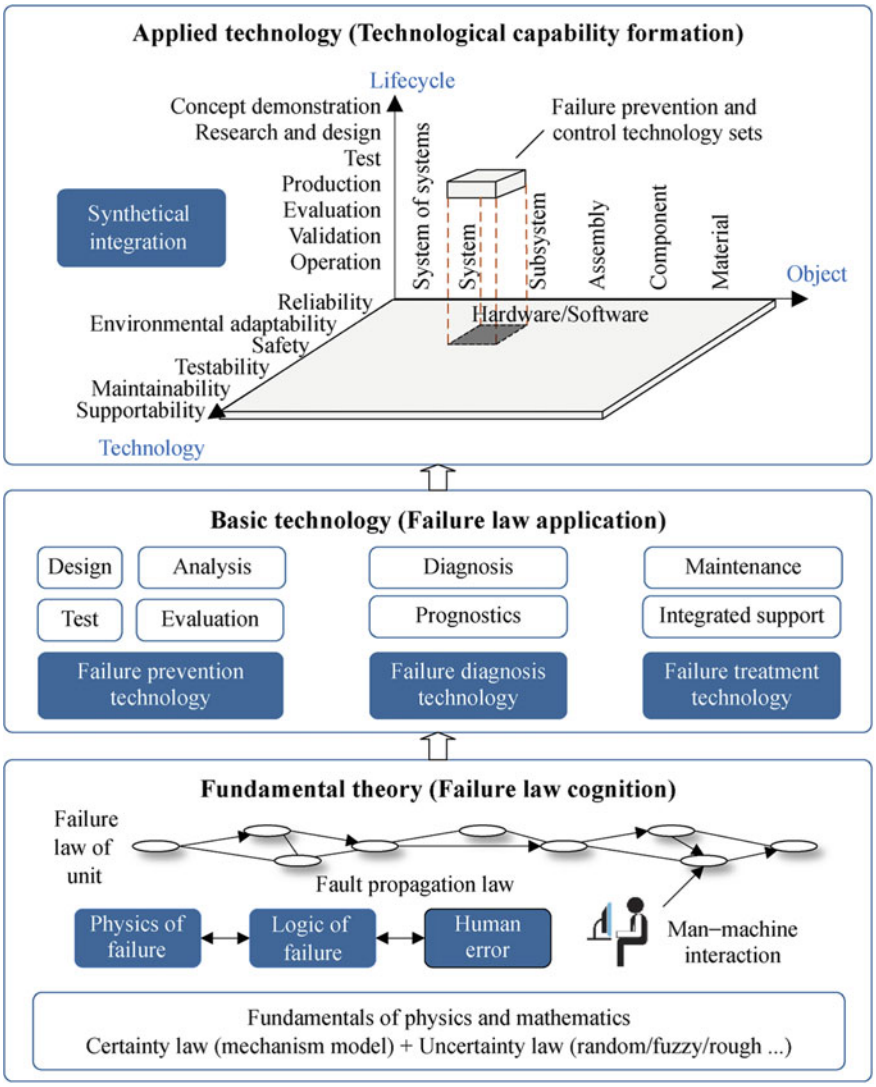


Fig. 2 Technological framework of RSE

2. Basic technology

On the basis of failure laws, a number of basic technologies for failure prevention, diagnosis, and treatment can be developed for RSE. The failure prevention technology is mainly related to the technology of failure prevention over the full lifecycle of a product, including the design, production, and use [29]. Existing redundancy technology, reduction technology, statistical process control technology, and reliability-centered maintenance (RCM) technology are all failure prevention technologies.

Failure diagnosis technology refers to the diagnosis and prediction of failures over the lifecycle of a product. Failure diagnosis focuses on the timely monitoring and isolation of failures, and it is concerned with the prediction of development trend and consequences of failures [30]. On the basis of failure diagnosis technology, the failure treatment technology refers to the technology for the timely and effective recovery of product functions when an uncontrollable failure occurs. It aims to repair the product, i.e., to restore the product function quickly, economically, and effectively, including the specific technology for repairing product failures, procedures for repairing product failures, and financing and supply of spare parts, tools, equipment, and personnel needed to repair product failures.

3. Applied technology

The application scope of RSE can be described in terms of three dimensions: Lifecycle, object, and technology. The lifecycle dimension represents the full process of systems engineering activities, including the concept demonstration, research and design, test, production, evaluation, validation, and operation. The object dimension refers to the physics items of all scales, including the materials, components, assemblies, subsystems, and systems. The technology dimension refers to the GQCs that can be applied by RSE, including the reliability, safety, maintainability, testability, supportability, and environmental adaptability. In particular, RSE can be applied as a failure prevention and control (P&C) technology set constructed via the integration of overall characterization methods, full lifecycle processes, and total system elements. Its core technology involves synthetic GQC integration, including synthetic GQC requirement determination with effectiveness simulations as the core, synthetic GQC design with unified function and failure models as the core, and synthetic operation and maintenance technology with PHM [31] as the core.

3 Conceptual and Operational Model of MBRSE

With the transformation of equipment development and application modes with technologies such as digital engineering and model-based system engineering technologies, the technical system of reliability system engineering has been gradually developed and promoted to modelization, digitalization, simulation and intelligence. Integrated with MBSE and RSE, the idea of MBRSE was first proposed by the authors of this chapter in 2016, and introduced in the 1st China Aviation Forum. In this section, the conceptual model and V-model-based operation mode of MBRSE are introduced.

3.1 Connotative, Features and Conceptual Models of MBRSE

On the basis of the product models in MBSE, a unified model system can be established by adding several other types of models such as environmental load model,

fault model, maintenance model, test model and support model. Furthermore, the GQC R&D demands are gradually obtained starting from the mission capability and support scenarios. Then based on the evolution of unified model, one could gradually identify and mitigate faults from the functional perspective to physical perspective, synchronously improve and optimize the GQC design, and verify the realization of the GQC R&D demands by simulations, as shown in Fig. 3.

MBRSE is an organic whole technique composed of models, methods, tools and environments, as shown in Fig. 4. Models are the key elements of MBRSE and also direct reflection of operation. The models in MBRSE include the product's functional performance models and failure P&C related models with dynamic evolution and traceable features. Methods refer to the specific technologies to achieve the product's functional performance and GQC characteristics. They are used to determine the implementation methods and processes of each task in the evolution of unified model. Tools are used to achieve all types of functional performance and GQC characteristics. They are mainly GQC related management and technological software, ensuring to improve the standardization, efficiency and effect of GQC work. Environment is an integration that organically combines models, methods, tools, resources, and manpower to achieve the GQC requirements of equipment under the support of environment.

Figure 5 shows the conceptual model of MBRSE. On the basis of usage demand, a comprehensive design issue is initially constructed. This comprehensive design can

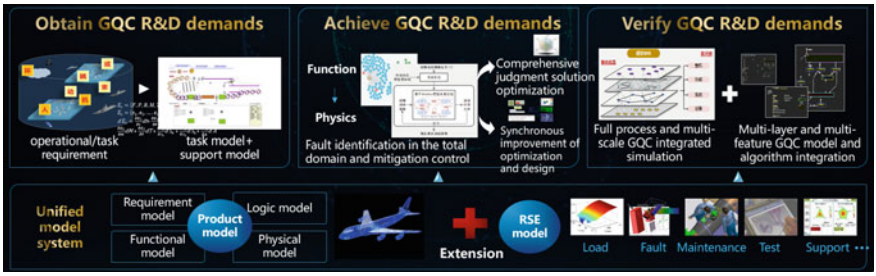


Fig. 3 Connotative meanings of MBRSE

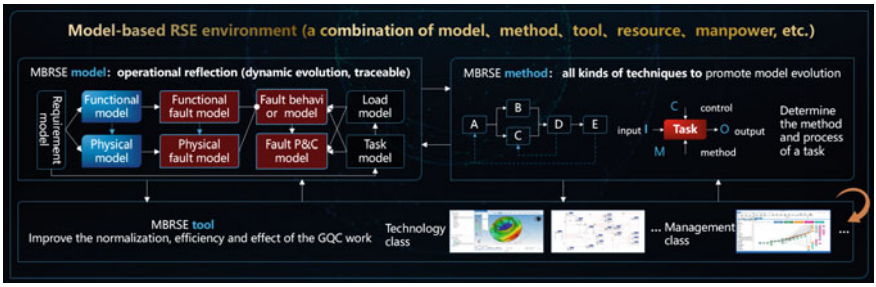


Fig. 4 Features of MBRSE

be decomposed into function and failure M&C designs. Comprehensive design issue can be analyzed and solved with various engineering methods. During the solution processing, the above-mentioned two types of designs should cooperate to reduce the number of design iterations. A failure M&C design is based on the cognition of failures and their control laws. With the ever-deepening cognition of design, the product design scheme is becoming increasingly in-depth and detailed and ranges from qualitative descriptions to quantitative calculations. The M&C process can be used to determine the failures and associated control laws. This understanding is based on the knowledge of the operation process/environment (load) that becomes increasingly clear with the advance of design processes. After solving all problems, system synthesis and evaluation are conducted to assess the solution process and verify the solution degree with regard to the comprehensive design issue. The above-mentioned process may require several iterations during practical product design until a satisfactory solution is reached with regard to the comprehensive design issue.

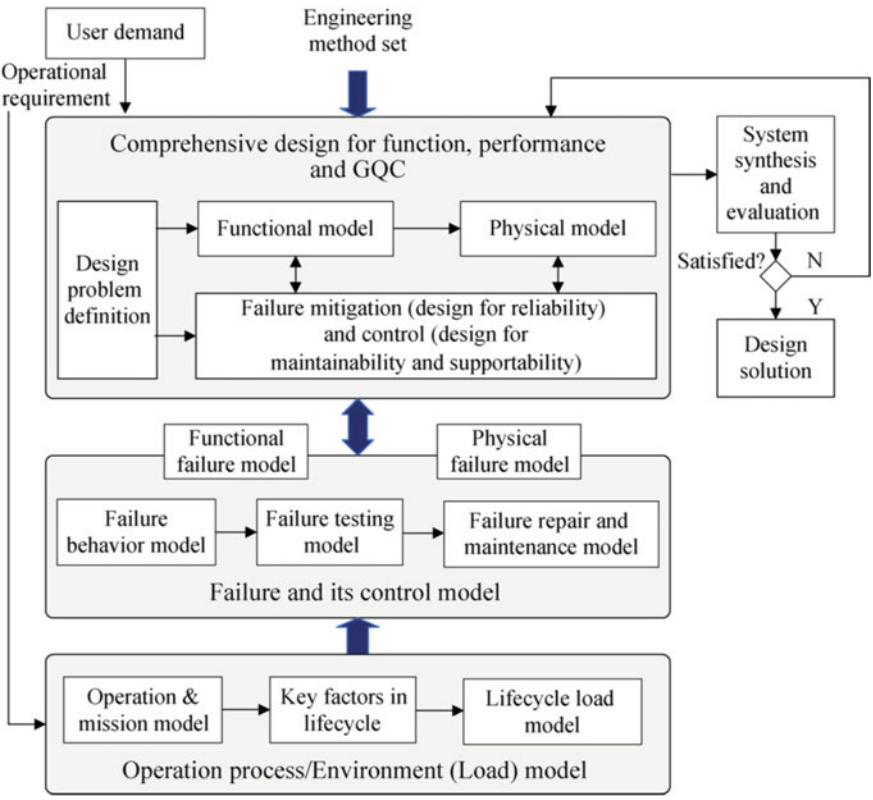


Fig. 5 Conceptual models of MBRSE

3.2 *Data and Information Sharing Mechanism and Process Control Mechanism of MBRSE*

The terminologies in GQC involve a large number of terms and concepts, with complex relationships with the product functional performance design concepts. Therefore, by using the ontology technology, this chapter provides a unified knowledge model of related concepts and their relationships, to achieve a unified expression of the functional performance and GQC data knowledge in the whole design process, and lay the foundation to establish a unified model for multiple engineering disciplines. There are four basic relationships to express integrated design ontology, namely, "part (P:)," "class (K:)," "instance," and "attribute (A:)." "Part" expresses the conceptual relationship between parts and the whole; "Class" expresses the inheritance relationship between concepts; "Instance" expresses the relationship between the instances of a concept and that concept; "Attribute" expresses that a concept is an attribute of another concept.

According to the relationships among the concepts of products, uses, functions, faults and their extended concepts, the established top-level ontology framework of the functional performance and GQC integrated design is shown in Fig. 6. The product structure consists of product meta, each with a generalized function, and various states. The fault is regarded as a state of the product meta. The functions and states of the product structure are not a simple sum of product meta. New functions and states may emerge, and they have to be expressed independently.

The design of MBRSE follows a bottom-up, iterative search for convergence process rule. It uses the process control mechanism of the integrated design to "precisely" control the entire process, to reduce the randomness in the search and solution process, and thus reduce the iterative numbers of the "design-feedback-redesign" cycles. Considering the inheritance problem of design meta, the meta-process is divided for brand-new product design and inherited product design.

1. Meta-process for brand-new product design meta

The new product meta has no prototype product to inherit; its design meta-process starts from the requirement model. During product design, on one hand, the design parameter x needs to be dynamically adjusted according to the design changes in the system; on the other hand, uncertain factors may change and propagate. Therefore, the product meta continues to experience the change process of "steady state-instability-steady state." The overall trend is that the product meta model is continuously refined, thus transformed from the initial functional model to the physical model, and then from the initial simple physical model to a detailed and precise physical model. Various methods and models based on product meta-models are constantly evolving, and the evaluation of various GQC design characteristics of products is also more accurate, thus continuously providing a direction for product design to achieve the design goals. A complete meta-process for the new product meta design is shown in Fig. 7.

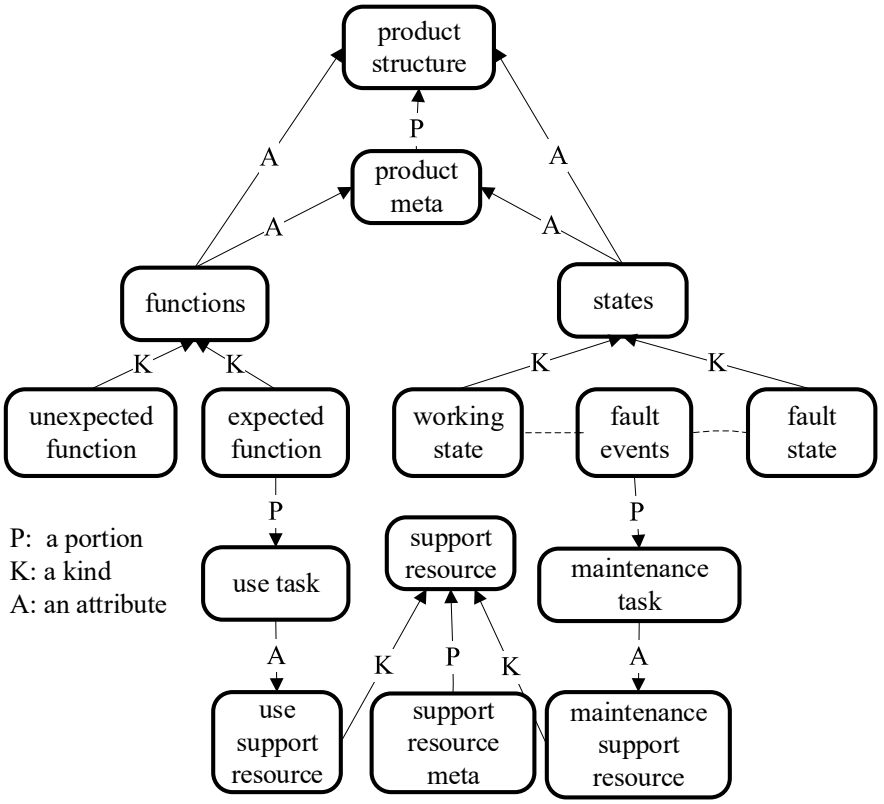


Fig. 6 Ontology framework of the integrated functional performance and GQC design

Moreover, during model evolution, it is necessary to establish relevant design analysis models according to all types of requirements, such as establishing reliability models of design elements based on reliability requirements (either statistical or PoF models), to build a model for different views of the same product. The goal of model transformation is to continuously obtain manufacturable products that meet various design requirements. The principle and basis of a model transformation are the engineering analysis models of these views.

2. Meta-process for inherited product design meta

In general product design, most design meta are completed by inheriting the existing design meta, and the re-design process is conducted using the similarity inheritance principle. The engineering process starts from the demand model, searches for possible matching design metas on the design requirements and loads from the design meta knowledge base, inherits design information according to their similarities with the source design meta, and generates an instance of the target design meta model. If the inherited design meta-template cannot be searched, the design process for a new product design meta-template is then started; otherwise, the design process for the

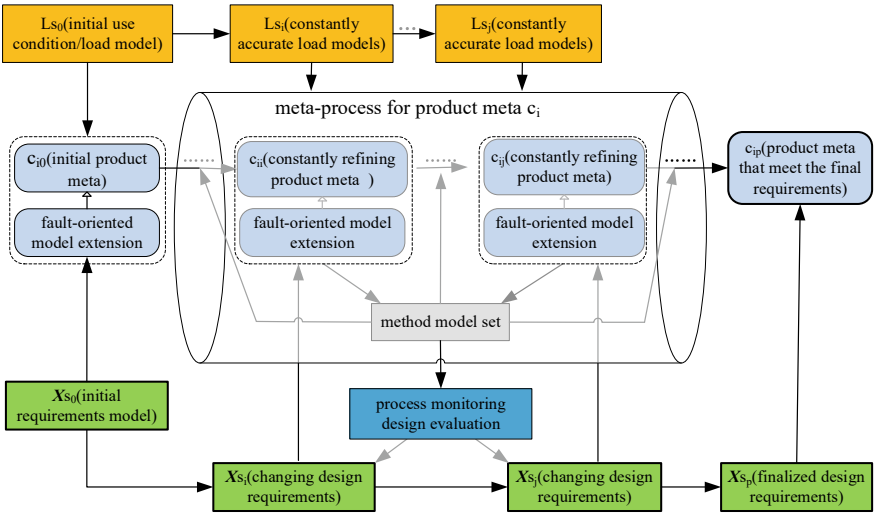


Fig. 7 Meta-process for the new product meta design

inherited design product meta-template is executed. During the design of the inherited product meta, the product meta model library is established by accumulating, summarizing and extracting historical experience. Moreover, the models of design method related to the product meta are also accumulated into a library, and those are connected to the product meta. During the refinement of design meta-model, it is necessary to continuously search the product meta-model by reusing the product meta-model and associated method model. Taking the advantages of the reuse of the model can reduce the number of model iterations in the process, and therefore shorten the development process. A complete meta-process for the new product meta design is shown in Fig. 8.

3.3 V-Model-Based Operation Model of MBRSE

The V-model-based MBRSE is innovatively proposed on the basis of the V-model of MBSE to organically integrate the function, performance, and GQC model into the MBSE process, as shown in Fig. 9. Driven by effectiveness, a complete model system is established by using the identification, mitigation, testing, and verification of failures as the core elements to achieve digitalized GQC engineering analysis throughout the entire product lifecycle. Synthetic function-GQC design and verification of multilevel products (up to the system of system (SOS) level) can be implemented by using approaches, such as multidimensional digital model coevolution, multitype failure simulation, and multithread closed-loop process management. Using the PHM technology, highly efficient and precise failure prognosis and

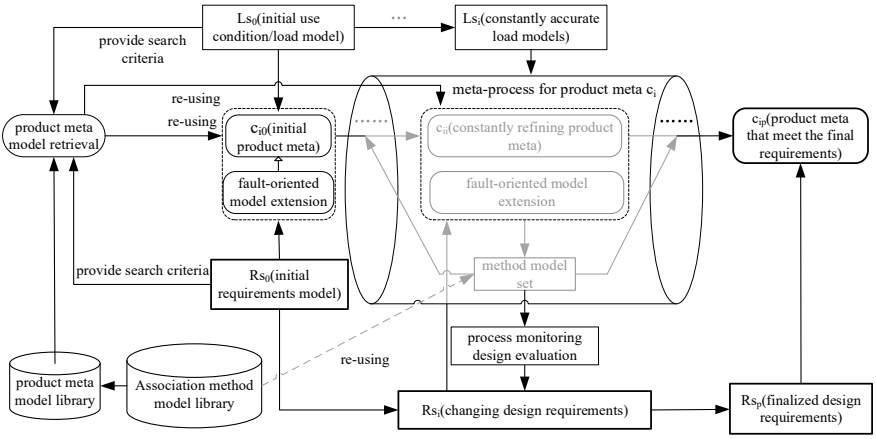


Fig. 8 Meta-process for inherited product design meta

prediction can be achieved to achieve digital and intelligent product maintenance. After the incorporation of GQC digital engineering into the systems engineering process, a forward GQC design process can be established. This process uses the GQC digital engineering V-model as the core and includes three-dimensional (3D) synthesis among performance–failure–health, multilevel product data transmission, and synthetic interaction between design analysis and simulation verification. The crucial technology required by the specific execution of V-model is shown in Sect. 4.

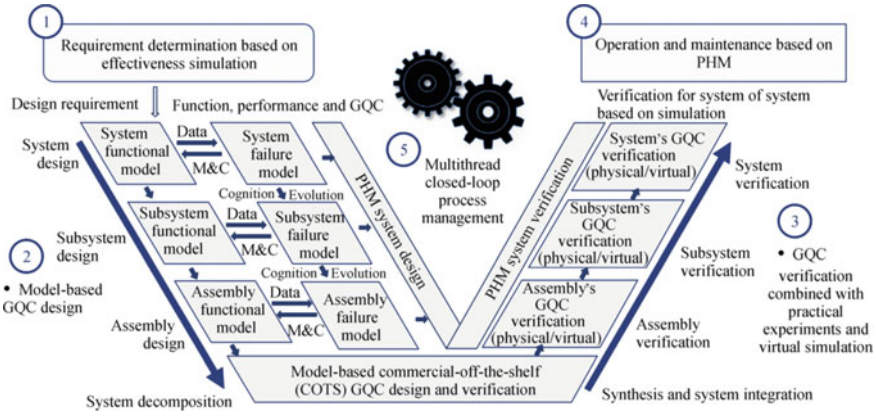


Fig. 9 V-model-based MBRSE

Fig. 10 Failure identification and mitigation in total domain during the MBRSE process

3.4 MBRSE Model Evolution Process

Based on the unified model evolution, the failure identification and mitigation in total domain can be accomplished during the MBRSE process, as shown in Fig. 10. In the requirement stage, the GQC requirements are obtained based on the effectiveness of simulation models. In the design stage, the functional and physical failure M&C requirements are obtained based on the functional and physical logic models, respectively. In addition, the integrated reliability design problems in multiple fields are solved based on the multidisciplinary mechanism model. Generally, the models used in the previous stage can provide inputs for the models in the next stage; moreover, the results obtained from the later stage can be used to verify the models in its previous stage model.

4 Crucial Technology for MBRSE Operation

4.1 Synthetic GQC Requirement Determination Based on Effectiveness Simulation

The top left side of the V-model indicates the efficiency-oriented, task-motivation-driven simulation verification techniques to obtain the GQC requirements. Originating from different tasks, the goal is to convert the equipment effectiveness demands into GQC requirements using simulation methods, providing accurate input for subsequent designs.

First, by considering the equipment characteristics such as task parallelism, element crossover, and human–machine cooperation, four types of elements, task, system, service, and environment, are abstracted as the agents. On this basis, a

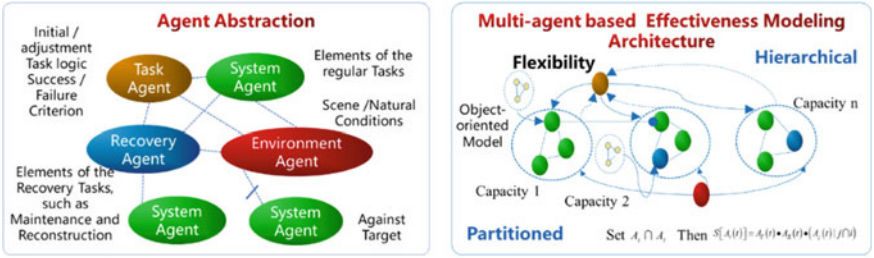


Fig. 11 An agent theory-based modeling and simulation framework

dynamic layering, partitioning, and flexibility effectiveness modeling and simulation framework can be established based on different agents, as shown in Fig. 11. Among them, the state of each agent is determined by its own state and the interactions with other agents.

Then, a dual-clock simulation mechanism is constructed to design a dynamic life clock for multilevel agents independent of the system simulation clock. On this basis, a life clock consumption based failure occurrence mechanism is constructed. In this way, when the life clock is cleared to zero, the agent turns into an unavailable state. Considering that the components and parts might become unavailable due to multiple factors such as degradation, random failure, and external shocks, three consumption mechanisms, uniform consumption, random consumption, and step consumption, were considered. For the life clock of high-level agents, the condition-based consumption mechanism can be established according to the failure propagation relationships among different elements.

Next, the maintenance safeguard behavior modeling mechanism is provided based on the agent action map and state transition. A unified maintenance event pool is built to manage the preventive maintenance, corrective maintenance, and condition-based maintenance events. By selecting the most appropriate agent, maintenance operations are dynamically generated by matching the maintenance personnel, facilities, equipment, and spare parts.

Then, an effectiveness evaluation technology is developed using Monte Carlo simulation and multidimensional performance criteria, and a batch of fault criteria under spatiotemporal dynamics can be built, including performance loss down to 0, space trajectory deviation, and standard coverage area. Then, after statistical analysis on all types of simulation process data of functional performance and maintenance, system status and changes, consumption status of the consumable resource, occupation status of the unconsumable resource, etc., the system effectiveness can be evaluated from the perspectives of performance, state, time, and resources.

Finally, parameter analysis, balance, comparison, and optimization technologies are developed based on the above-mentioned highly precise effectiveness simulations. These technologies can quickly support the simulation determination of GQC requirements combined with the batch processing of similar entity attributes.

To summarize, the proposed synthetic GQC requirement determination technique using effectiveness simulations can overcome the limitations of traditional GQC requirement demonstration methods, such as the similarity methods and empirical methods, and also provide an essential foundation for the independent development and innovation of novel equipment.

4.2 Model-Driven Comprehensive GQC Design

The left side of V-model indicates the model-driven comprehensive GQC design approach, aiming to allocate the reliability requirements obtained from effectiveness simulations to the different product levels in a scaled-down sequence and simultaneously acquire a corresponding digital design plan [27, 32]. With the continuous development of MBSE concept, the GQC design technique has been proposed with unified failure modeling and mitigation control as the core, as shown in Fig. 12.

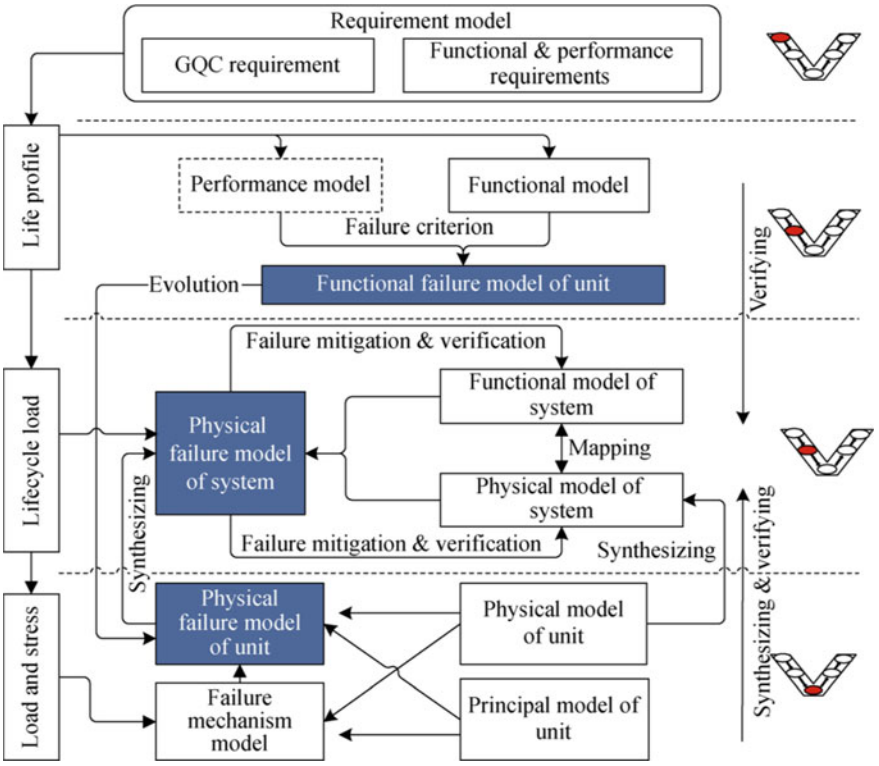


Fig. 12 GQC design based on failure models and its control

1. Identification and mitigation of functional failures

The functional failures can be determined from the perspective of the disappearance or reduction in the ability to achieve and maintain functional requirements, as shown in Fig. 13. Assuming that product level i contains n_i function achievement requirements F_{Rij} (j = 1, 2, …, n_i), it can be easily found that each function achievement requirement has multiple states {normal, function loss, discontinuity, incompleteness, offset, …}. When the function achievement requirements lay in the states of discontinuity, incompleteness, offset, etc., it means that the ability to maintain the function achievement requirements continues to decline, indicating that a fault will occur at this time. The potential failure modes of functions can then be gradually identified by establishing a functional failure clue sheet. Meanwhile, a functional failure model can be established by considering the transfer relationships among the effects of these functional failures [33, 34]. Association sets of the key physical failure modes can be determined by comprehensively considering the occurrence probabilities and consequences of these failure modes and the increase in the M&C status of relevant failures caused by the mitigation of an individual failure mode. Effective implementation of the corresponding improvement and compensation measures should be ensured using a closed-loop mitigation control technology.

2. M&C of physical failures

On the basis of the above-mentioned functional failure models and mapping relationships between the functionalities and physical models, physical failure modes can be systematically identified according to unmitigated functional failures considering physicochemical processes, device/raw material/component characteristics, temperature/vibration, and other internal and external loads [25], as shown in Fig. 14. The root causes of the physical failure originate from the physicochemical effects of the physical unit, as well as the effect of internal and external loads. In addition to the

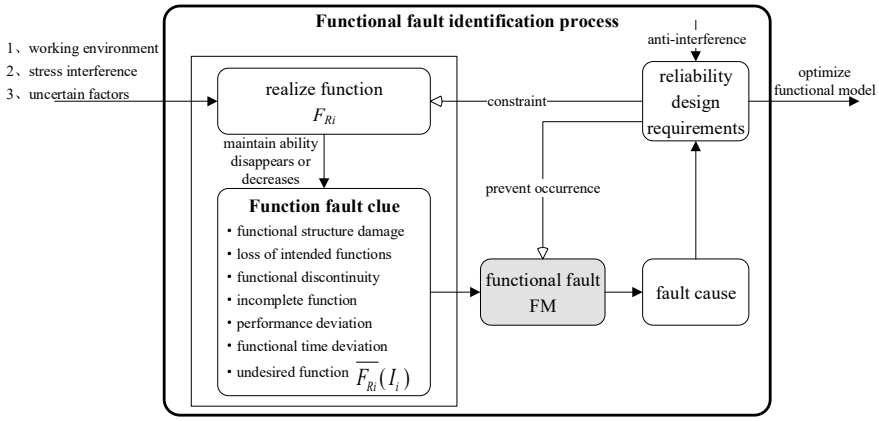


Fig. 13 Functional fault identification process

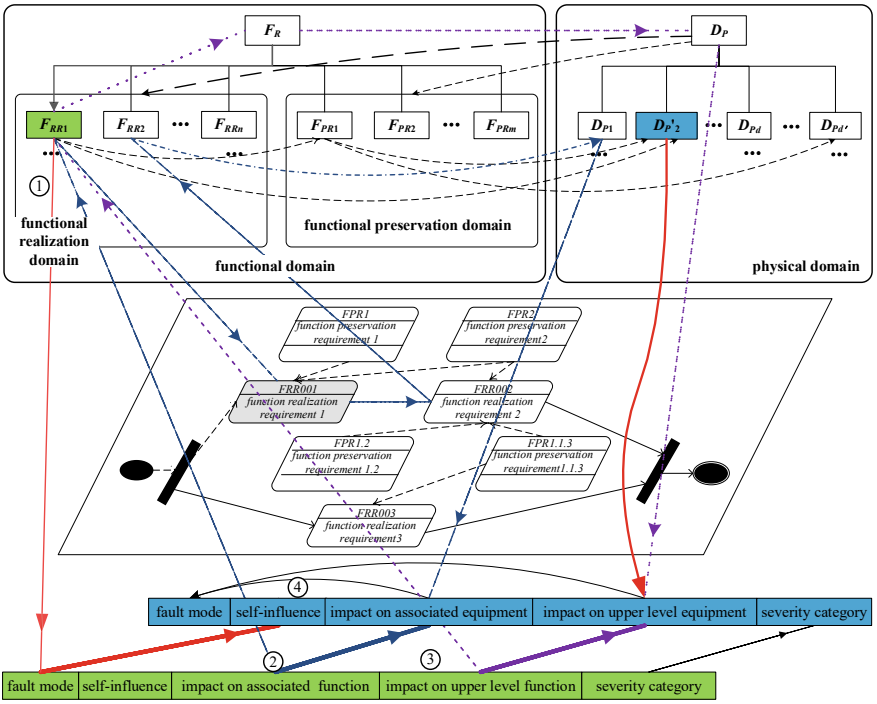


Fig. 14 Identification of physical failure modes

above-mentioned function-physics mapping relationship, the following conditions can be further combined to identify the physical faults that cause the relevant fault impacts:

1. The working principle of the physical components, including the component structure and materials, information and energy in its physicochemical action process
2. The characteristics of the devices, raw materials, and mechanical parts to construct the physical component, such as the tensile strength, flexural strength, compressive strength, seismic strength, expansion coefficient, density, dielectric constant of the metals, temperature resistance, and antimagnetic properties of electronic materials; For components/raw materials/parts, according to the above-mentioned conditions, PoF models can be constructed for fault analysis and identification.
3. The internal and external loads applied on the physical components, including vibration, temperature, humidity, electromagnetic, pressure, mold, salt spray, sand and dust, and remainder particles.

Next, under the premise of a clear understanding of the failure mechanisms of physical units such as mechanical devices, electronics, and software, association sets

of the key physical failure modes can be determined by comprehensively considering the occurrence probability and consequences of these failure modes and an increase in the M&C status of relevant failures due to the mitigation of an individual failure mode [35]. PoF-based mitigation mechanisms are introduced to achieve closed-loop mitigation control for determining and optimizing the design parameters of the corresponding physical units, mitigating underlying failure causes, avoiding failure occurrence, and preventing physical units from being controlled by these failures.

3. **M&C of coupled failures**

In the synthetic process of a system, system-level failures can be identified by comprehensively considering the interface, transmission, error propagation, and potential functionality failures. First, the interface relationships in between the i product meta are sorted out, and the interface failure modes are analyzed to obtain the interface failure mode models. Then, by analyzing the load conditions over the life cycle of the system, together with the structure of the system-level product, layout of the product meta, and internal loads in the system, the local loads of each product meta can be obtained. According to these local loads, the possible new failure modes of the product meta, and the functional impacts of the new failure modes on the product meta and system are further analyzed. Next, by comprehensively analyzing the functional failure and physical failure modes of each product meta in the system, both the functional and the physical failure coupled models can be established. Furthermore, combined with the interface failure mode models, the set of new failure modes emerging from system integration can be obtained. This failure mode set, together with the system failure modes induced by product meta-faults, composes the complete set of system failure modes, as shown in Fig. 15. These system-level failures are usually regarded as coupled failures and must be mitigated and controlled with their relevant failures [36].

4.3 Multilevel GQC Verification Combined with Practical Experiments and Virtual Simulations

During the integration of a given product, integrated verification should be conducted via a scaled-up sequence, starting from the assembly level, via the subsystem and system levels, to the SOS level, as shown in Fig. 16. At each level, certain tests and weak link analysis steps are conducted to verify the identified failures and defects and to determine new failures and defects. The experimental data acquired from these tests can be further implemented to verify the product GQCs and effectiveness at different levels. The approach combining physical tests with virtual simulations has been widely applied for verification due to a notable advancement in simulation technology.

1. Verification at the assembly level has usually been conducted via multistress synthesis GQC tests (for instance Temperature/Humidity/Vibration Combined

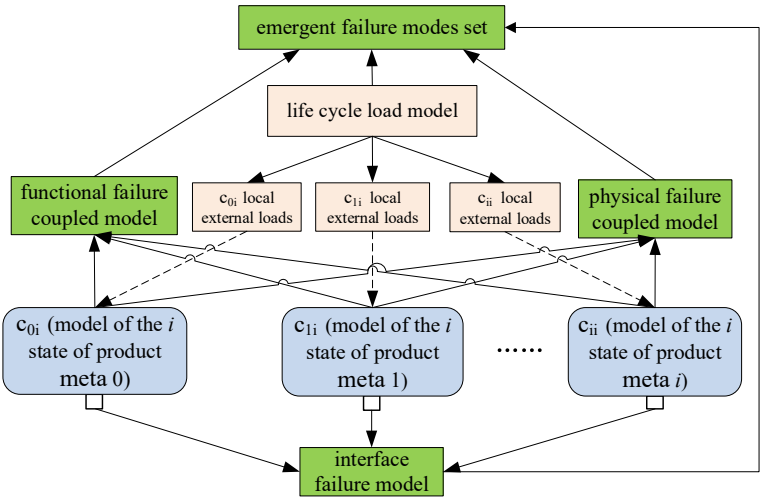


Fig. 15 Integration of product meta-fault rules into product structure

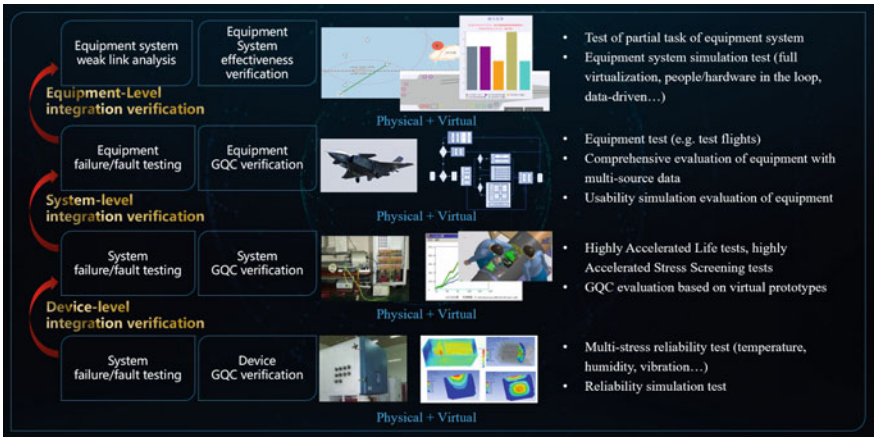


Fig. 16 Multilayer GQC verification tests

Environmental Test, Temperature/Humidity/Vibration/Low Pressure Combined Environmental Test, etc.) on a physical product, and GQC simulations (for instance Finite Element Analysis, Computational Fluid Dynamics, and PoF) on its virtual representation.

2. Verification at the subsystem level has usually been conducted via Highly Accelerated Life tests, Highly Accelerated Stress Screening tests, and GQC evaluation simulations based on system models (for instance Reliability Block Diagram, Bayesian Network, and Artificial Neural Network).

3. Verification at the system level has usually been conducted via product-level full-scale tests (for instance, Outfield Flight Test, and Full-Scale Simulated Laboratory Test of Comprehensive Environment), multisource data-driven virtual GQC evaluation simulations, and availability simulations.
4. Verification at the SOS level has usually been conducted via real-task tests, virtual effectiveness simulations (via different types such as all simulation, humans-in-the-loop, hardware-in-the-loop, and data driven, complementary to the models and technologies discussed in Sect. 4.1).

In recent years, the U.S. has promoted digital engineering technology in defense procurement. Digital test qualification plays a central role in digital engineering, and it must be integrated with digital engineering activities. During a typical "design-test-validation" cycle in the product development process, the physical model is used to provide the initial design of system performance. The designed system performance is tested in various environmental conditions and the test results are then fed back into the modeling process to validate or update the physical model. The verified model forms the basis for improving the system design or related requirements on modeling and simulation validation. In the future, the test data should be extensively collected from all the stages in the MBRSE process to validate the GQC characteristics, and a "model-test-validation" cycle method should be used to continuously evaluate, calibrate, and improve the product at all the hierarchical levels.

4.4 Operation and Maintenance Using PHM as the Core During Operation

The top right side of V-model indicates the product operation and maintenance using PHM technology as the core. As a notable development and supplement to the current reliability engineering field, this method focuses on the usage phase of the product via an organic integration of GQC. Its concrete implementation includes development, operation, and maintenance stages [37]. In the development stage, PHM system design and verification are needed, including construction of the PHM index system, establishment of the system configuration with total elements, breakthrough of the key failure detection and prognosis techniques in the space, time and symptom dimensions, and completion of the PHM system on the basis of a variety of tools and methods [30, 31]. In this process, the development of PHM system should be suitably coordinated with the design of product function, performance, and GQCs. At the operation and maintenance phase, the health status of a product containing a PHM system can be improved or maintained at a high level through failure prediction to plan reasonable maintenance tasks [38] and support resources and advance scheduling on the basis of the concepts of autonomous assurance, task effects, and health status. The maintenance cost can be reduced. The main work tasks include the following:

1. In development planning of PHM technology: Chinese industries in aviation, aerospace, shipbuilding, and industrial manufacturing need a step-by-step technology development plan, such as: from the fundamental studies on PoF and mechanism models on components and parts, to the subsystem-level product monitoring, diagnosis, prediction and fault-tolerant control technology studies, and then to the design and development of PHM systems for the product and equipment.
2. In the basic theories of PHM technology: (1) New intelligent sensor studies should be carried out for failure diagnosis and prediction. (2) Accurate measurement of system health status should be carried out to determine the health evolution trend of the system, on different hierarchical levels from components, products to the system. (3) Improvement of the effectiveness and accuracy of prediction should be studied by considering the uncertainty of failure prediction. (4) Studies of maintenance decision optimization, health self-healing, and control should be effectively achieved for the health management of complex systems.
3. In engineering application of the PHM technology: (1) Application studies of system integration should be carried out to achieve the PHM collaborative design, by using the parallel engineering principles to synchronize PHM and design of the product. (2) Studies should be focused on the cognition of the diagnosis and prediction ability, functional simulation test verification, and quantitative performance evaluation.

4.5 Multithread Closed-Loop Process Management

Many GQC tasks involve the interaction of people, data, activities, and resources that require effective management. Considering the practical application, three threads should be conducted as follows:

1. The first thread is conducted on the basis of the allocation, prediction, and quantitative evaluation of GQCs, reflecting the kernel realization of user requirements for developers, where various quantitative requirements are subject to hard design constraints.
2. The second thread is conducted on the basis of the implementation and conformance inspection of qualitative GQC criteria. This process reflects the accumulation and reuse of the GQC design experience of the developer that may effectively improve the GQC level of similar products.
3. The third thread is conducted on the basis of the closed-loop mitigation of failures. This process is mainly applied to identify, eliminate, and control the consequences of new failures caused by new product principles, processes, materials, and system integration methods.

5 MBRSE Platform and Reliability Digital Twin (RDT)

5.1 MBRSE Platform

On the basis of the fundamental theory and approaches of MBRSE, the Institute of Reliability Engineering at Beihang University developed the model-based 4th generation of the MBRSE platform by innovating key technologies for visualizing multilayer and multidimensional GQC data, multidimensional failure data analysis, flexible process instruction chains, full-scale failure recognition, closed-loop M&C, reliability knowledge mapping construction, and mining. This MBRSE platform is developed by taking the product, failure and environmental models as the core; Fig. 17 shows its entire structure. It includes more than 10 model-driven GQC design software tools to greatly reduce the number of reliability work items and improve work effectiveness. It also integrates a basic GQC knowledge system with storage, mining, and intelligent push functions to further improve the GQC knowledge reuse ability, and a dynamic visual monitoring and decision-making system to promote a transparent design process of RMS. Furthermore, by integrating a digital development environment and unifying technology and management synthesis, the platform is capable to achieve a novel model-driven form that combines RMS data, process, and characteristics to reach an international leading level.

5.2 RDT

RDT is the latest technology carrier of MBRSE combing the advantages of both the MBRSE technology and digital twin (DT) technology. It changes the traditional reliability technology that used to be "soft," and leads to the revolution and evolution of the reliability technology. The concept of DT first appeared as "virtual digital representation equivalent to a physical product" proposed by Professor Grieves in 2003 [39]. In 2014, the release of the DT White Paper established its core architecture of "physical space, virtual space and interconnection" [39]. Since its birth, DT has had a natural relationship to the reliability technology. For instance, it pays special attention on the monitoring and prediction of product failures and health status. The U.S. Department of Defense (DoD) has used the DT technology very early to solve problems such as the health maintenance of spacecrafts [40].

RDT is not only the enhanced perception of product reliability and health status using the DT technology, but also enriches and completes DT owing to the reliability technology. The reliability and DT exhibit a complementary and mutually promoting relationship to each other, as shown in Fig. 18. On one hand, digital intelligence analysis, evaluation, and control can be achieved based on digital entities and twin data. On the other hand, high fidelity, high precision, and quasi real-time feedback of reliability and health status of reliability and health status can also be provided by using the reliability technology. The RDT has the following features and advantages:

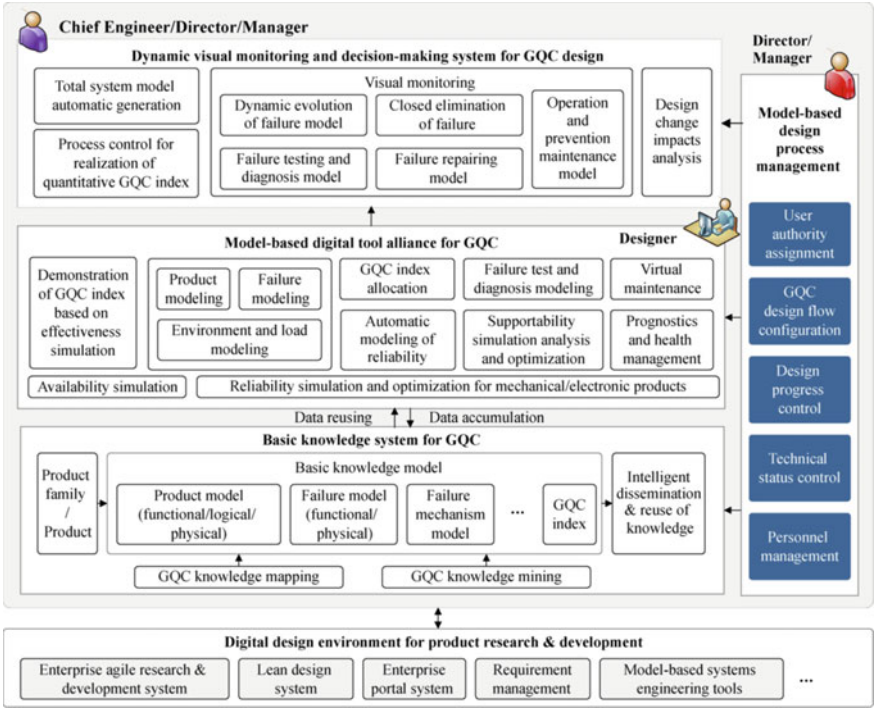


Fig. 17 MBRSE Platform

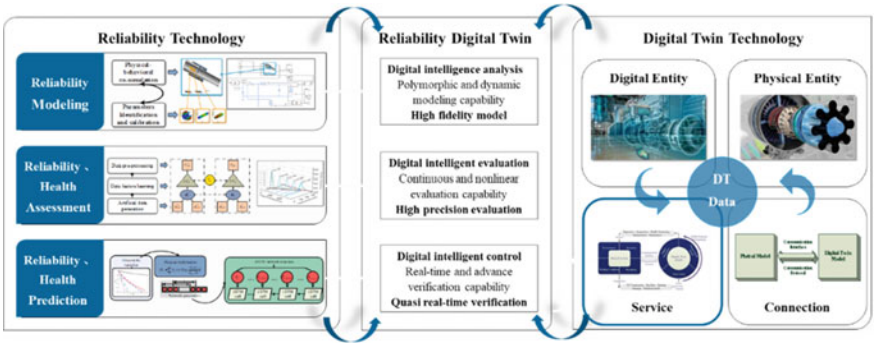


Fig. 18 Relationship between the reliability and DT technology

- Generation of virtual entities from physical entities: precise mapping from the physical space to virtual space.
- Control of physical entities with virtual entities: implementation of virtual entities to control physical entities.

- Interaction between physical and virtual entities: personalized and precise intelligent operation and maintenance on virtual entities to enhance physical entities.
- Mutual existence of physical and virtual entities: physical and virtual entities interact and empower symbiosis and coprosperity.

RDT needs to be built and implemented in the design stage of the product and then used and evolved in the operation phase. It covers the full product lifecycle, including the design, test, manufacturing, operation, and maintenance. Its object covers all the hierarchical levels of the product, such as materials, components, equipment, subsystems, and systems. In the design stage, it is necessary to build a digital model that is equivalent to the physical entity considering reliability and other GQCs, combined with the functional and performance design of the product. The model should also have the ability to interface or map with all types of monitoring, test, and operational data during the use of the product, and achieve the simultaneous delivery of both the RDT model and physical entity of the product. In the operation stage, RDT will serve as an enabling tool for product users, to provide high-fidelity, quasi real-time reliability/health status information for the operation and maintenance of the product, and thus to provide accurate perception of the system health status and dynamic operation and maintenance decision-making based on the status of individual units. Furthermore, the RDT model that has been already built and run can iteratively feedback more integrated, comprehensive, and real information to generate the reliability characteristics required in the design of new products.

The key RDT technologies mainly include the following: development of the MBRSE based RDT technology framework, establishment of high-fidelity RDT models, bidirectional mapping of reliability characteristics in between physical entities and DTs, and real-time update and evolution of RDT models. At present, the RDT technology is still in the early stages of development, and faces many challenges in the implementation. However, its future development potential and great benefits have begun to reach a wide consensus. The Institute of Reliability Engineering at Beihang University first designed and developed a brake system RDT test platform in China (as shown in Fig. 19) that can be used for RDT-related technology research and teaching experiments and can also provide valuable practical experience for RDT development of other types of systems.

6 Concluding Remarks and Outlook

This chapter reviews the history of development of RSE in China over the past 30 years. The fundamental theories and technologies of RSE have experienced a typical development process from statistics-based methods to PoF-based methods, and they now occur at a new level that emphasizes collaboration between mathematics and physics in addition to integrated optimization. RSE technology has followed the development direction on interdisciplinary and professional integration and has

Fig. 19 Brake system RDT test platform

currently entered the health engineering stage. In particular, equal attention has been paid to "good birth" and "healthcare conditions," and failures have been used as the core elements. Health has been used as the goal. Prevention, diagnosis, and treatment have been used as the approaches, and the synthetic design of performance and GQCs and a PHM-integrated platform have been used as the support.

The technological framework, conceptual and operational models, crucial technologies, and methodology of MBRSE are emphatically introduced. Combined with PHM, MBRSE has transformed the original concept from RCM to intelligent prognosis and health management by using health as the goal. This approach has been transformed from the pursuit of the perfect stage without failure to the allowance of disease to a certain extent while ensuring health that occurs more often in practical situations.

Representative directions include but are not limited to the following.

1. Cross-scale-based synthetic GQC design focusing on the macroscopic effectiveness, microscale failure mechanisms, and intelligent design processes. SOS level: To develop effectiveness simulation analysis and design optimization methods of intelligent SOS. System and subsystem level: To study intelligent GQC design by providing a preliminary model design, intelligent failure identification, and mitigation. Component and part level: Synthesis-based design technology considering multiphysics, multiperspectives, new processes, and new materials.
2. Cognitive computing-based health assessment, diagnosis, and prognosis techniques. These techniques may improve the perception, cognition, and capability of failure prediction throughout the full lifecycle of equipment.
3. Government-industry data exchange program in China. A vast amount of raw data can be retrieved by mining with a certain deposition quality; higher profits can be obtained through data exchange. This condition can be achieved by establishing a

quality information exchange platform jointly organized by the military, government, and industry with regular/real-time interactive engineering data, failure experience data, reliability and maintainability data, and measurement data.

In the future, RSE technologies and platform will be continuously innovated and promoted in the military and civil fields by taking the MBRSE platform as the carrier through a variety of means such as trail promotion, comprehensive promotion, and upgrade promotion. During these processes, the RSE technologies should also be improved from the perspectives of normalization, quantification, and optimization abilities.

References

1. Kuo W (2015) Risk and reliability are part of our life. In: Proceedings of 1st international conference on reliability systems engineering (ICRSE). IEEE Press, Beijing, pp 1–6
2. Yang SL, Wang JM, Shi LY, Tan YJ, Qiao F (2018) Engineering management for high-end equipment intelligent manufacturing. Front Eng Manage 5(4):420–450
3. Yu Y (2019) Low-carbon technology calls for comprehensive electricity-market redesign. Front Eng Manage 6(1):128–130
4. Si SB, Zhao JB, Cai ZQ, Dui HY (2020) Recent advances in system reliability optimization driven by importance measures. Front Eng Manage 7(3):335–358
5. Oestern HJ, Huels B, Quirini W, Pohlemann T (2000) Facts about the disaster at Eschede. J Orthop Trauma 14(4):287–290
6. Smith MS (2003) NASA's space shuttle Columbia: synopsis of the report of the Columbia Accident Investigation Board. Congressional Research Service, The Library of Congress, Washington, DC
7. Cusumano MA (2021) Boeing's 737 MAX: a failure of management, not just technology. Commun ACM 64(1):22–25
8. Blanchard BS, Fabrycky WJ (1990) Systems engineering and analysis, 4th edn. Prentice Hall, Englewood Cliffs, NJ
9. Sohlenius G (1992) Concurrent engineering. CIRP Ann 41(2):645–655
10. Kaâniche M, Laprie JC, Blanquart JP (2000) Dependability engineering of complex computing systems. In: Proceedings of 6th International Conference on Engineering of Complex Computer Systems September. IEEE Press, Tokyo, pp 36–46
11. Thornton AC, Donnelly S, Ertan B (2000) More than just robust design: Why product development organizations still contend with variation and its impact on quality. Res Eng Design 12(3):127–143
12. Yang WM, Ruan L, Tu QC (1995) Reliability systems engineering—Theory and practice. Acta Aeronautica et Astronautica Sinica 16(S1):1–8 (in Chinese)
13. Yang WM (1995) Overview of reliability, maintainability and supportability. National Defense Industry Press, Beijing (in Chinese)
14. Kang R, Wang ZL (2005) Framework of theory and technique about reliability systems engineering. Acta Aeronautica et Astronautica Sinica 26(5):633–636 (in Chinese)
15. Kang R, Wang ZL (2007) Overview of quality management of the full system, full characteristics, and full process of product. Technol Foundation Natl Defense 4:25–29 (in Chinese)
16. Shi SY (2007) Chinese Military Encyclopedia—General introduction to military technology. Encyclopedia of China Publishing House, Beijing (in Chinese)
17. Ren Y, Wang ZL, Yang DZ, Feng Q, Sun B (2021) Model-based reliability systems engineering. National Defense Industry Press, Beijing (in Chinese)

18. Li Y, Sun B, Wang ZL, Ren Y (2017) Ontology-based environmental effectiveness knowledge application system for optimal reliability design. J Comput Inf Sci Eng 17(1):011005
19. Zhao Q, Jia X, Cheng ZJ, Guo B (2018) Bayes estimation of residual life by fusing multisource information. Front Eng Manage 5(4):524–532
20. Wu ZY, Wang ZL, Feng Q, Sun B, Qian C, Ren Y, Jiang X (2018) A gamma process-based prognostics method for CCT shift of high power white LEDs. IEEE Trans Electron Devices 65(7):2909–2916
21. Xia Q, Wang ZL, Ren Y, Sun B, Yang D, Feng Q (2018) A reliability design method for a lithium-ion battery pack considering the thermal disequilibrium in electric vehicles. J Power Sources 386:10–20
22. Fan DM, Wang ZL, Liu LL, Ren Y (2016) A modified GO-FLOW methodology with common cause failure based on discrete time Bayesian network. Nucl Eng Des 305:476–488
23. Ren Y, Fan DM, Ma XR, Wang ZL, Feng Q, Yang DZ (2018) A GOFLOW and dynamic Bayesian network combination approach for reliability evaluation with uncertainty: a case study on a Nuclear Power Plant. IEEE Access 6:7177–7189
24. Yang X, Sun B, Wang ZL, Qian C, Ren Y, Yang DZ, Feng Q (2018) An alternative lifetime model for white light emitting diodes under thermal-electrical stresses. Materials 11(5):817–828
25. Qian C, Sun ZQ, Fan JJ, Ren Y, Sun B, Feng Q, Yang DZ, Wang ZL (2020) Characterization and reconstruction for stochastically distributed void morphology in nano-silver sintered joints. Mater Des 196:109079
26. Wang Z, Kang R, Xie LY (2009) Dynamic reliability modeling of systems with common cause failure under random load. Maintenance Reliab 43(3):47–54
27. Yang DZ, Ren Y, Wang ZL, Liu LL, Sun B (2015) A novel logic-based approach for failure modes mitigation control and quantitative system reliability analyses. Maintenance Reliab 17(1):100–106
28. Che HY, Zeng SK, Guo JB (2019) Reliability assessment of man–machine systems subject to mutually dependent machine degradation and human errors. Reliab Eng Syst Saf 190:106504
29. Yang DZ, Ren Y, Wang ZL, Xiao J (2014) Design-in of reliability through axiomatic design. J Beijing Univ Aeronaut Astronaut 40(1):63–68 (in Chinese)
30. Tian Y, Ma J, Lu C, Wang ZL (2015) Rolling bearing fault diagnosis under variable conditions using LMD-SVD and extreme learning machine. Mech Mach Theory 90:175–186
31. Wang ZL, Cui YQ, Shi JY (2017) A framework of discrete-event simulation modeling for prognostics and health management (PHM) in airline industry. IEEE Syst J 11(4):2227–2238
32. Yang DZ, Ren Y, Wang ZL, Liu LL (2012) Decision-making of failure modes mitigation program considering coupling relationship among failure modes. J Beijing Univ Aeronaut Astronaut 38(10):1389–1394 (in Chinese)
33. Li ZF, Ren Y, Liu LL, Wang ZL (2015) Parallel algorithm for finding modules of large-scale coherent fault trees. Microelectron Reliab 55(9–10):1400–1403
34. Ren Y, Fan DM, Wang ZL, Yang DZ, Feng Q, Sun B, Liu LL (2018) System dynamic behavior modeling based on extended GO methodology. IEEE Access 6:22513–22523
35. Sun B, Pan WY, Wang ZL, Yung KC (2015) Envelope probability and EFAST-based sensitivity analysis method for electronic prognostic uncertainty quantification. Microelectron Reliab 55(9–10):1384–1390
36. Liu LL, Fan DM, Wang ZL, Yang DZ, Cui JJ, Ma XR, Ren Y (2019) Enhanced GO methodology to support failure mode, effects and criticality analysis. J Intell Manuf 30(3):1451–1468
37. Li YR, Peng SZ, Li YT, Jiang W (2020) A review of condition-based maintenance: Its prognostic and operational aspects. Front Eng Manage 7(3):323–334
38. Wu QL, Feng Q, Ren Y, Xia Q, Wang ZL, Cai BP (2021) An intelligent preventive maintenance method based on reinforcement learning for battery energy storage systems. IEEE Trans Ind Inform. https://doi.org/10.1109/TII.2021.3066257. (in press)
39. Grieves M (2014) Digital twin: manufacturing excellence through virtual factory replication. In: Digital twin white paper
40. Glaessgen E, Stargel D (2012) The digital twin paradigm for future NASA and U.S. air force vehicles. In: 53rd structures, structural dynamics, and materials conference. Honolulu, Hawaii, pp 1–14

Professor Hong-Zhong Huang's Selected Journal Papers

Year of 2023

[1] Zeng Y, Huang T, Li YF, Huang HZ (2023) Reliability modeling for power converter in satellite considering periodic phased mission. Reliab Eng Syst Saf 232:109039

[2] Qian HM, Huang HZ, Wei J (2023) Time-variant reliability analysis for a complex system based on active-learning Kriging model. ASCE-ASME J Risk Uncert Eng Syst Part A: Civil Engi 9(1):04022055

Year of 2022

[1] Liu Z, Liu X, Huang HZ, Zhu P, Liang Z (2022) A new inherent reliability modeling and analysis method based on imprecise Dirichlet model for machine tool spindle. Ann Oper Res 311(1):295–310

[2] Zhou J, Huang HZ, Li YF, Guo J (2022) A framework for fatigue reliability analysis of high-pressure turbine blades. Ann Oper Res 311(1):489–505

[3] Mi J, Lu N, Li YF, Huang HZ, Bai L (2022) An evidential network-based hierarchical method for system reliability analysis with common cause failures and mixed uncertainties. Reliab Eng Syst Saf 220:108295

[4] Shi Y, Lu Z, Huang H, Liu Y, Li Y, Zio E, Zhou Y (2022) A new preventive maintenance strategy optimization model considering lifecycle safety. Reliab Eng Syst Saf 221:108325

[5] Li XY, Xiong X, Guo J, Huang HZ, Li X (2022) Reliability assessment of non-repairable multi-state phased mission systems with backup missions. Reliab Eng Syst Saf 223:108462

[6] Xiahou T, Zeng Z, Liu Y, Huang HZ (2021) Measuring conflicts of multisource imprecise information in multistate system reliability assessment. IEEE Trans Reliab 71(4):1417–1434

[7] Guo J, Li YF, Peng W, Huang HZ (2022) Bayesian information fusion method for reliability analysis with failure-time data and degradation data. Qual Reliab Eng Int 38(4):1944–1956

Y. Liu et al. (eds.), *Advances in Reliability and Maintainability Methods and Engineering Applications*, Springer Series in Reliability Engineering,
https://doi.org/10.1007/978-3-031-28859-3

[8] Chen Y, Huang HZ Rao Y, He Z Lai P, Chen Y, … Liu C (2022) Degradation assessment of 1.2-kV SiC MOSFETs and comparative study with 1.2-kV Si IGBTs under power cycling. Microelectron Reliab 132:114528
[9] Hu JM, Huang HZ, Li YF, Gao HY (2022) Bayesian prior information fusion for power law process via evidence theory. Commun Stat-Theory Methods 51(14):4921–4939
[10] Qian HM, Huang HZ, Li YF, Wei J (2022) A novel approach for multi-output structural system reliability problem with small failure probability. Int J Numer Eng 123(23):5986–5999
[11] Bai S, Li YF, Huang HZ, Ma Q, Lu N (2022) A probabilistic combined high and low cycle fatigue life prediction framework for the turbine shaft with random geometric parameters. Int J Fatigue 165:107218

Year of 2021

[1] Qian HM, Li YF, Huang HZ (2021) Time-variant system reliability analysis method for a small failure probability problem. Reliab Eng Syst Saf 205:107261
[2] Li XY, Huang HZ, Li, YF, Xiong X (2021) A Markov regenerative process model for phased mission systems under internal degradation and external shocks. Reliab Eng Syst Saf 215:107796
[3] Liu Y, Zhang Q, Ouyang Z, Huang HZ (2021) Integrated production planning and preventive maintenance scheduling for synchronized parallel machines. Reliab Eng Syst Saf 215:107869
[4] Li X, Li YF, Li H, Huang HZ (2021) An algorithm of discrete-time Bayesian network for reliability analysis of multilevel system with warm spare gate. Qual Reliab Eng Int 37(3):1116–1134
[5] Huang T, Xiahou T, Li YF, Qian HM, Liu Y, Huang HZ (2021) Reliability assessment of wind turbine generators by fuzzy universal generating function. Eksploatacja i Niezawodność-Maintenance Reliab 23(2):308–314
[6] Bai S, Li YF, Huang HZ, Yu A, Zeng Y (2021) An improved petri net for fault analysis of an electronic system with hybrid fault of software and hardware. Eng Failure Anal 120:105077
[7] Hu J, Huang HZ, Li YF (2021) Bayesian estimation of a power law process with incomplete data. J Syst Eng Electron 32(1):243–251
[8] Huang P, Huang HZ, Li YF, Qian HM (2021) An efficient and robust structural reliability analysis method with mixed variables based on hybrid conjugate gradient direction. Int J Numer Methods Eng 122(8):1990–2004
[9] Qian HM, Huang T, Huang HZ (2021) A single-loop strategy for time-variant system reliability analysis under multiple failure modes. Mech Syst Signal Process 148:107159
[10] Huang CG, Huang, HZ, Li YF, Peng W (2021) A novel deep convolutional neural network-bootstrap integrated method for RUL prediction of rolling bearing. J Manufact Syst 61:757–772

]11] Huang P, Huang HZ, Li YF, Li H (2021) Positioning accuracy reliability analysis of industrial robots based on differential kinematics and saddlepoint approximation. Mech Mach Theory 162:104367

Year of 2020

[1] Levitin G, Finkelstein M, Huang HZ (2020) Optimal mission abort policies for multistate systems. Reliab Eng Syst Saf 193:106671
[2] Li XY, Li YF, Huang HZ (2020) Redundancy allocation problem of phased-mission system with non-exponential components and mixed redundancy strategy. Reliab Eng Syst Saf 199:106903
[3] Qian HM, Li YF, Huang HZ (2020) Time-variant reliability analysis for industrial robot RV reducer under multiple failure modes using Kriging model. Reliab Eng Syst Saf 199:106936
[4] Huang CG, Yin X, Huang HZ Li, YF (2019) An enhanced deep learning-based fusion prognostic method for RUL prediction. IEEE Trans Reliab 69(3):1097–1109
[5] Li YF, Liu Y, Huang T, Huang HZ, Mi J (2020) Reliability assessment for systems suffering common cause failure based on Bayesian networks and proportional hazards model. Qual Reliab Eng Int 36(7):2509–2520
[6] Qian HM, Li YF, Huang HZ (2020) Improved model for computing time-variant reliability based on outcrossing rate. ASCE-ASME J Risk Uncert Eng Syst Part A: Civil Eng 6(4):04020043
[7] Qian, HM, Huang T, Huang HZ (2020) Time-dependent reliability model of structures under different stochastic loads. ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Eng 6(2):021003
[8] Huang P, Huang HZ, Huang T, Qian HM (2020) A framework for structural reliability analysis based on conjugate sensitivity factor and saddlepoint approximation. J Mech Sci Technol 34(9):3617–3627
[9] Yu A, Huang HZ, Li H, Li YF, Bai S (2020) Reliability analysis of rolling bearings considering internal clearance. J Mech Sci Technol 34:3963–3971
[10] Li H, Soares CG, Huang HZ (2020) Reliability analysis of a floating offshore wind turbine using Bayesian Networks. Ocean Eng 217:107827

Year of 2019

[1] Levitin G, Xing L, Huang HZ (2019) Dynamic availability and performance deficiency of common bus systems with imperfectly repairable components. Reliab Eng Syst Saf 189:58–66
[2] Levitin G, Finkelstein M, Huang HZ (2019) Scheduling of imperfect inspections for reliability critical systems with shock-driven defects and delayed failures. Reliab Eng Syst Saf 189:89–98
[3] Levitin G, Xing L, Huang HZ (2019) Security of separated data in cloud systems with competing attack detection and data theft processes. Risk Anal 39(4):846–858

[4] Zeng Y, Li YF, Li XY, Huang HZ (2019) Tolerance-based reliability and optimization design of switched-mode power supply. Qual Reliab Eng Int 35(8):2774–2784
[5] Guo J, Huang HZ, Peng W, Zhou J (2019) Bayesian information fusion for degradation analysis of deteriorating products with individual heterogeneity. Proc Inst Mech Eng Part O: J Risk Reliab 233(4):615–622
[6] Huang CG, Huang HZ, Li YF (2019) A bidirectional LSTM prognostics method under multiple operational conditions. IEEE Trans Ind Electron 66(11):8792–8802
[7] Hu JM, Huang HZ, Li YF (2019) Reliability growth planning based on information gap decision theory. Mech Syst Signal Process 133:106274
[8] Levitin G, Xing L, Huang HZ (2019) Cost effective scheduling of imperfect inspections in systems with hidden failures and rescue possibility. Appl Math Model 68:662–674
[9] Qian HM, Huang HZ, Li YF (2019) A novel single-loop procedure for time-variant reliability analysis based on Kriging model. Appl Math Model 75:735–748

Year of 2018

[1] Mi J, Li YF, Peng W, Huang HZ (2018) Reliability analysis of complex multi-state system with common cause failure based on evidential networks. Reliab Eng Syst Saf 174:71–81
[2] Li XY, Huang HZ, Li YF (2018) Reliability analysis of phased mission system with non-exponential and partially repairable components. Reliab Eng Syst Saf 175:119–127
[3] Li XY, Li YF, Huang HZ, Zio E (2018) Reliability assessment of phased-mission systems under random shocks. Reliab Eng Syst Saf 180:352–361
[4] Zhang X, Gao H, Huang HZ, Li YF, Mi J (2018) Dynamic reliability modeling for system analysis under complex load. Reliab Eng Syst Saf 180: 345–351
[5] Peng W, Balakrishnan N, Huang HZ (2018) Reliability modelling and assessment of a heterogeneously repaired system with partially relevant recurrence data. Appl Math Model 59:696–712
[6] Li XY, Huang HZ Li, YF, Zio E (2018) Reliability assessment of multi-state phased mission system with non-repairable multi-state components. Appl Math Model 61:181–199
[7] Li H, Huang HZ, Li YF, Zhou J, Mi J (2018) Physics of failure-based reliability prediction of turbine blades using multi-source information fusion. Appl Soft Comput 72:624–635
[8] Zheng B, Huang HZ, Guo W Li YF, Mi J (2018) Fault diagnosis method based on supervised particle swarm optimization classification algorithm. Intel Data Anal 22(1):191–210
[9] Zheng B Li YF, Huang HZ (2018) Intelligent fault recognition strategy based on adaptive optimized multiple centers. Mech Syst Signal Process 106, 526–536

[10] Zhang X, Gao H, Huang HZ (2018) Total fatigue life prediction for welded joints based on initial and equivalent crack size determination. Int J Damage Mech 27(7):1084–1104
[11] Guo J, Li YF, Zheng B, Huang HZ (2018) Bayesian degradation assessment of CNC machine tools considering unit non-homogeneity. J Mech Sci Technol 32:2479–2485
[12] Zuo FJ, Li YF, Huang HZ (2018) Reliability analysis for fatigue damage of railway welded bogies using Bayesian update based inspection. Smart Struct Syst 22(2):193–200

Year of 2017

[1] Wang HK, Li YF, Huang HZ, Jin T (2017) Near-extreme system condition and near-extreme remaining useful time for a group of products. Reliab Eng Syst Saf 162:103–110
[2] Peng W, Li YF, Yang YJ, Mi J, Huang HZ (2017) Bayesian degradation analysis with inverse Gaussian process models under time-varying degradation rates. IEEE Trans Reliab 66(1):84–96
[3] Liu Z, Yu L, Li YF, Mi J, Huang HZ (2017) Comparisons of two non-probabilistic structural reliability analysis methods for aero-engine turbine disk. Int J Turbo Jet-Engines 34(3):295–303
[4] Zhou J, Huang HZ, Peng Z (2017) Fatigue life prediction of turbine blades based on a modified equivalent strain model. J Mech Sci Technol 31:4203–4213
[5] Zhong B, Chen XH, Pan R, Wang J, Huang HZ, Deng WH, Wang ZZ, Xie RQ, Liao DF (2017) The effect of tool wear on the removal characteristics in high-efficiency bonnet polishing. Int J Adv Manuf Technol 91:3653–3662
[6] Zhu SP, Lei Q, Huang HZ, Yang YJ, Peng W (2017) Mean stress effect correction in strain energy-based fatigue life prediction of metals. Int J Damage Mech 26(8):1219–1241

Year of 2016

[1] Mi J, Li YF, Yang YJ, Peng W, Huang HZ (2016) Reliability assessment of complex electromechanical systems under epistemic uncertainty. Reliab Eng Syst Saf 152:1–15
[2] Peng W, Li YF, Mi J, Yu L, Huang HZ (2016) Reliability of complex systems under dynamic conditions: A Bayesian multivariate degradation perspective. Reliab Eng Syst Saf 153:75–87
[3] Zhu SP, Huang HZ, Peng W, Wang HK, Mahadevan S (2016) Probabilistic physics of failure-based framework for fatigue life prediction of aircraft gas turbine discs under uncertainty. Reliab Eng Syst Saf 146:1–12
[4] Wang HK, Huang HZ, Li YF, Yang YJ (2015) Condition-based maintenance with scheduling threshold and maintenance threshold. IEEE Trans Reliab 65(2):513–524
[5] Peng W, Li YF, Yang YJ, Zhu SP, Huang, HZ (2016) Bivariate analysis of incomplete degradation observations based on inverse Gaussian processes and copulas. IEEE Trans Reliab 65(2):624–639

[6] Yang YJ, Peng W, Zhu SP, Huang HZ (2016) A Bayesian approach for sealing failure analysis considering the non-competing relationship of multiple degradation processes. Eksploatacja i Niezawodność-Maintenance Reliab 18(1):10–15
[7] Yin YC, Huang HZ, PeNG W, Li YF, Mi J (2016) An E-Bayesian method for reliability analysis of exponentially distributed products with zero-failure data. Eksploatacja i Niezawodność-Maintenance Reliab 18(3):445–449
[8] Xie C, Wang P, Wang Z, Huang H (2016) Corrosion reliability analysis considering the coupled effect of mechanical stresses. ASCE-ASME J Risk Uncertain Eng Syst Part B: Mech Eng 2(3):031001
[9] Wang, Z, Zhang, X, Huang, HZ, Mourelatos, ZP (2016) A simulation method to estimate two types of time-varying failure rate of dynamic systems. Journal of Mechanical Design, 138(12)
[10] Peng Z, Huang HZ, Wang HK Zhu SP, Lv Z (2016) A new approach to the investigation of load interaction effects and its application in residual fatigue life prediction. Int J Damage Mech 25(5):672–690

Year of 2015

[1] Moghaddass R, Zuo MJ, Liu Y, Huang HZ (2015) Predictive analytics using a nonhomogeneous semi-Markov model and inspection data. IIE Trans 47(5):505–520
[2] Liu Y, Lin P, Li YF, Huang HZ (2014) Bayesian reliability and performance assessment for multi-state systems. IEEE Trans Reliab 64(1):394–409
[3] Liu Y, Zuo MJ, Li YF, Huang HZ (2015) Dynamic reliability assessment for multi-state systems utilizing system-level inspection data. IEEE Trans Reliab 64(4):1287–1299
[4] Mi J, Li YF, Liu Y, Yang YJ, Huang HZ (2015) Belief universal generating function analysis of multi-state systems under epistemic uncertainty and common cause failures. IEEE Trans Reliab 64(4):1300–1309
[5] Peng W, Li YF, Yang YJ, Mi J, Huang HZ (2015) Leveraging degradation testing and condition monitoring for field reliability analysis with time-varying operating missions. IEEE Trans Reliab 64(4):1367–1382
[6] Wang HK, Li YF, Liu Y, Yang YJ, Huang HZ (2015) Remaining useful life estimation under degradation and shock damage. Proc Inst Mech Eng Part O: J Risk Reliab 229(3):200–208
[7] Li YF, Mi J, Liu YU, Yang YJ, Huang HZ (2015) Dynamic fault tree analysis based on continuous-time Bayesian networks under fuzzy numbers. Proc Inst Mech Eng Part O: J Risk Reliab 229(6):530–541
[8] Yang JP, Huang HZ, Liu Y, Li YF (2015) Quantification classification algorithm of multiple sources of evidence. Int J Inf Technol Decis Mak 14(05):1017–1034
[9] Meng D, Li, YF, Huang HZ, Wang Z, Liu Y (2015) Reliability-based multidisciplinary design optimization using subset simulation analysis and its application in the hydraulic transmission mechanism design. J Mech Des 137(5):051402

[10] Zhu SP, Huang HZ, Li Y, Liu Y, Yang Y (2015) Probabilistic modeling of damage accumulation for time-dependent fatigue reliability analysis of railway axle steels. Proc Inst Mech Eng Part F: J Rail Rapid Transit 229(1):23–33
[11] Lv Z, Huang HZ, Zhu SP, Gao H, Zuo F (2015) A modified nonlinear fatigue damage accumulation model. Int J Damage Mech 24(2):168–181
[12] Yuan R, Li H, Huang HZ Zhu SP, Gao H (2015) A nonlinear fatigue damage accumulation model considering strength degradation and its applications to fatigue reliability analysis. Int J Damage Mech 24(5):646–662
[13] Zuo FJ, Huang HZ, Zhu SP Lv Z, Gao H (2015) Fatigue life prediction under variable amplitude loading using a non-linear damage accumulation model. Int J Damage Mech 24(5):767–784
[14] Huang HZ, Wang HK, Li YF, Zhang L, Liu Z (2015) Support vector machine based estimation of remaining useful life: current research status and future trends. J Mech Sci Technol 29:151–163
[15] Li YF, Zhu SP, Li J, Peng W, Huang HZ (2015) Uncertainty analysis in fatigue life prediction of gas turbine blades using Bayesian inference. Int J Turbo Jet-Engines 32(4):319–324

Year of 2014

[1] Liu Y, Wang W, Huang HZ, Li Y, Yang Y (2014) A new simulation model for assessing aircraft emergency evacuation considering passenger physical characteristics. Reliab Eng Syst Saf 121, 187–197
[2] Peng W, Li YF, Yang YJ, Huang HZ, Zuo MJ (2014) Inverse Gaussian process models for degradation analysis: A Bayesian perspective. Reliab Eng Syst Saf 130, 175–189
[3] Liu Z, Li YF, He LP, Yang YJ, Huang HZ (2014) A new fault tree analysis approach based on imprecise reliability model. Proc Inst Mech Eng Part O: J Risk Reliab 228(4):371–381
[4] Xiao NC, Li YF, Yu L, Wang Z, Huang HZ (2014) Saddlepoint approximation-based reliability analysis method for structural systems with parameter uncertainties. Proc Inst Mech Eng Part O: J Risk Reliab 228(5):529–540
[5] Yang YJ, Peng W, Meng D, Zhu SP, Huang HZ (2014) Reliability analysis of direct drive electrohydraulic servo valves based on a wear degradation process and individual differences. Proc Inst Mech Eng Part O: J Risk Reliab 228(6):621–630
[6] Xiao NC, Li YF, Yang Y, Yu L, Huang HZ (2014) A novel reliability method for structural systems with truncated random variables. Struct Saf 50:57–65
[7] Peng W, Liu Y, Li YF, Zhu SP, Huang HZ (2014) A Bayesian optimal design for degradation tests based on the inverse Gaussian process. J Mech Sci Technol 28:3937–3946
[8] Meng D, Huang HZ, Wang Z, Xiao NC, Zhang XL (2014) Mean-value first-order saddlepoint approximation based collaborative optimization for multidisciplinary problems under aleatory uncertainty. J Mech Sci Technol 28:3925–3935

[9] Huang HZ, Cui PL, Peng W, Gao HY, Wang HK (2014) Fatigue lifetime assessment of aircraft engine disc via multi-source information fusion. Int J Turbo Jet-Engines 31(2):167–174

Year of 2013

[1] Peng W, Huang HZ, Li Y, Zuo MJ, Xie M (2013) Life cycle reliability assessment of new products—A Bayesian model updating approach. Reliab Eng Syst Saf 112, 109–119
[2] Liu Y, Huang HZ, Wang Z, Li Y, Yang Y (2013) A joint redundancy and imperfect maintenance strategy optimization for multi-state systems. IEEE Trans Reliab 62(2):368–378
[3] Peng W, Huang HZ, Xie M, Yang Y, Liu Y (2013) A Bayesian approach for system reliability analysis with multilevel pass-fail, lifetime and degradation data sets. IEEE Trans Reliab 62(3):689–699
[4] Xiao NC, Huang HZ, Li YF, Wang Z, Zhang XL (2013) Non-probabilistic reliability sensitivity analysis of the model of structural systems with interval variables whose state of dependence is determined by constraints. Proc Inst Mech Eng Part O: J Risk Reliab 227(5):491–498
[5] Jin T, Yu Y, Huang HZ (2013) A multiphase decision model for system reliability growth with latent failures. IEEE Trans Syst Man Cybern: Syst 43(4):958–966
[6] Liu Y, Huang HZ, Ling D (2013) Reliability prediction for evolutionary product in the conceptual design phase using neural network-based fuzzy synthetic assessment. Int J Syst Sci 44(3):545–555
[7] Zhu SP, Huang HZ, Smith R, Ontiveros V, He LP, Modarres M (2013) Bayesian framework for probabilistic low cycle fatigue life prediction and uncertainty modeling of aircraft turbine disk alloys. Probab Eng Mech 34:114–122
[8] Huang HZ, Zhang X, Meng DB, Wang Z, Liu Y (2013) An efficient approach to reliability-based design optimization within the enhanced sequential optimization and reliability assessment framework. J Mech Sci Technol 27:1781–1789
[9] He LP, Huang HZ Pang Y, Li Y, Liu Y (2013) Importance identification for fault trees based on possibilistic information measurements. J Intell Fuzzy Syst 25(4):1013–1026
[10] Prabhu Gaonkar RS, Xie M, Huang HZ (2013) Optimizing maritime travel time reliability. Proc Inst Mech Eng Part M: J Eng Maritime Environ 227(2):167–176

Year of 2012

[1] Liu Y, Huang HZ, Zhang X (2011) A data-driven approach to selecting imperfect maintenance models. IEEE Trans Reliab 61(1):101–112
[2] Huang HZ, Zhang H, Li Y (2012) A new ordering method of basic events in fault tree analysis. Qual Reliab Eng Int 28(3):297–305

[3] Li YF, Huang HZ, Liu Y, Xiao N, Li H (2012) A new fault tree analysis method, fuzzy dynamic fault tree analysis. Eksploatacja i Niezawodnosc-Maintenance Reliab 14:208–214
[4] Huang HZ (2012) Structural reliability analysis using fuzzy sets theory. Eksploatacja i Niezawodność-Maintenance Reliab 14, 284–294
[5] Xiao NC, Huang HZ, Wang Z, Li Y, Liu Y (2012) Reliability analysis of series systems with multiple failure modes under epistemic and aleatory uncertainties. Proc Inst Mech Eng Part O: J Risk Reliab 226(3):295–304
[6] Zhang X, Huang HZ, Xu H (2010) Multidisciplinary design optimization with discrete and continuous variables of various uncertainties. Struct Multidisc Optim 42:605–618
[7] Huang HZ, Zhang X, Liu Y, Meng D, Wang Z (2012) Enhanced sequential optimization and reliability assessment for reliability-based design optimization. J Mech Sci Technol 26:2039–2043
[8] Wang Z, Li G, Huang HZ, Zhang XL, Li Y (2012) Reliability-based design optimization for box-booms considering strength degradation and random total working time. J Mech Sci Technol 26:2045–2049
[9] Wang Z, Huang HZ, Li Y, Pang Y, Xiao NC (2012) An approach to system reliability analysis with fuzzy random variables. Mech Mach Theory 52:35–46
[10] Xiao NC, Huang HZ, Wang Z, Liu Y, Zhang XL (2012) Unified uncertainty analysis by the mean value first order saddlepoint approximation. Struct Multidisc Optim 46:803–812
[11] Zhu SP, Huang HZ, Ontiveros V, He LP, Modarres M (2012) Probabilistic low cycle fatigue life prediction using an energy-based damage parameter and accounting for model uncertainty. Int J Damage Mech 21(8):1128–1153
[12] Huang HZ, Gong J, Zuo MJ, Zhu SP, Liao Q (2012) Fatigue life estimation of an aircaft engine under different load spectrums. Int J Turbo Jet-Engines 29(4):259–267

Year of 2011

[1] Guo S, Huang HZ, Wang Z, Xie M (2011) Grid service reliability modeling and optimal task scheduling considering fault recovery. IEEE Trans Reliab 60(1):263–274
[2] Wang Z, Huang HZ, Li Y, Xiao NC (2011) An approach to reliability assessment under degradation and shock process. IEEE Trans Reliab 60(4):852–863
[3] Pang Y, Huang HZ, He L, Wang Z, Xiao NC (2011) Convex sublattice based reliability theory. Eksploatacja i Niezawodność-Maintenance Reliab (3):56–61
[4] Huang HZ, Wang ZL, Li YF, Huang B, Xiao NC, He LP (2011) A nonprobabilistic set model of structural reliability based on satisfaction degree of interval. Mechanics 17(1):85–92
[5] Huang HZ, Zhang X, Yuan W, Meng D, Zhang X (2011) Collaborative reliability analysis under the environment of multidisciplinary design optimization. Concurr Eng 19(3):245–254

[6] Xiao N, Huang HZ, Li Y, He L, Jin T (2011) Multiple failure modes analysis and weighted risk priority number evaluation in FMEA. Eng Fail Anal 18(4):1162–1170
[7] Yang J, Huang HZ, He LP, Zhu SP, Wen D (2011) Risk evaluation in failure mode and effects analysis of aircraft turbine rotor blades using Dempster–Shafer evidence theory under uncertainty. Eng Failure Anal 18(8):2084–2092
[8] Xiao NC, Huang HZ, Wang Z, Pang Y, He L (2011) Reliability sensitivity analysis for structural systems in interval probability form. Struct Multidisc Optim 44:691–705
[9] Wang ZL, Huang HZ, Du L (2011) Competitive failure reliability analysis under fuzzy degradation data. Appl Soft Comput 11(3):2964–2973
[10] Liu Y, Huang HZ (2011) Optimal replacement policy for fuzzy multi-state element. J Multiple Valued Log Soft Comput 17(1):69–92
[11] Liu Y, Chen W, Arendt P, Huang HZ (2011) Toward a better understanding of model validation metrics. J Mech Des 133(7):071005
[12] Qu J, Liu Z, Zuo MJ, Huang HZ (2011) Feature selection for damage degree classification of planetary gearboxes using support vector machine. In: Proc Inst Mech Eng Part C: J Mech Eng Sci 225(9):2250–2264

Year of 2010

[1] Liu Y, Li Y, Huang HZ, Zuo MJ, Sun Z (2010) Optimal preventive maintenance policy under fuzzy Bayesian reliability assessment environments. IIE Trans 42(10):734–745
[2] Liu Y, Huang HZ (2010) Optimal selective maintenance strategy for multi-state systems under imperfect maintenance. IEEE Trans Reliab 59(2):356–367
[3] Liu Y, Huang HZ (2010) Optimal replacement policy for multi-state system under imperfect maintenance. IEEE Trans Reliab 59(3):483–495
[4] Wu J, Ng TSA, Xie M, Huang HZ (2010) Analysis of maintenance policies for finite life-cycle multi-state systems. Comput Ind Eng 59(4):638–646
[5] Liu Y, Huang HZ (2010) Reliability assessment for fuzzy multi-state systems. Int J Syst Sci 41(4):365–379
[6] Wang Z, Huang HZ, Du X (2010) Optimal design accounting for reliability, maintenance, and warranty. J Mech Des 132(1):011007
[7] Liu Y, Yin X, Arendt P, Chen W, Huang HZ (2010) A hierarchical statistical sensitivity analysis method for multilevel systems with shared variables. J Mech Des 132(3):031006
[8] Wang Z, Huang HZ, Liu Y (2010) A unified framework for integrated optimization under uncertainty. J Mech Des Trans ASME 132(5):051008
[9] Zhang X, Huang HZ (2010) Sequential optimization and reliability assessment for multidisciplinary design optimization under aleatory and epistemic uncertainties. Struct Multidisc Optim 40:165–175
[10] Zhang X, Huang HZ, Xu H (2010) Multidisciplinary design optimization with discrete and continuous variables of various uncertainties. Struct Multidisc Optim 42, 605–618

[11] Huang HZ, Yu H, Zhang X, Zeng S, Wang Z (2010) Collaborative optimization with inverse reliability for multidisciplinary systems uncertainty analysis. Eng Optim 42(8):763–773
[12] Zhang X, Zhang XL, Huang HZ, Wang Z, Zeng S (2010) Possibility-based multidisciplinary design optimization in the framework of sequential optimization and reliability assessment. Int J Innov Comput Inf Control 6(11):5287–5297
[13] Wang D, Miao Q, Fan X, Huang HZ (2009) Rolling element bearing fault detection using an improved combination of Hilbert and wavelet transforms. J Mech Sci Technol 23:3292–3301

Year of 2009

[1] Huang HZ, Qu J, Zuo MJ (2009) Genetic-algorithm-based optimal apportionment of reliability and redundancy under multiple objectives. IIE Trans 41(4):287–298
[2] Liu Y, Huang HZ (2009) Comment on "A framework to practical predictive maintenance modeling for multi-state systems" by Tan CM and Raghavan N. [Reliab Eng Syst Saf 2008; 93(8):1138–50]. Reliab Eng Syst Saf 94(3):776–780
[3] Huang HZ, An ZW (2008) A discrete stress-strength interference model with stress dependent strength. IEEE Trans Reliab 58(1):118–122
[4] An ZW, Huang HZ Lin D (2009) An approach to reliability evaluation of multiple V-belt drives considering the deviation of belt length. Proc Inst Mech Eng Part O: J Risk Reliab 223(2):159–166
[5] Huang HZ, Zhang X (2009) Design optimization with discrete and continuous variables of aleatory and epistemic uncertainties. J Mech Des Trans ASME 131(3):031006
[6] Wang GB Huang HZ, Liu Y, Zhang X, Wang Z (2009) Uncertainty estimation of reliability redundancy in complex systems based on the Cross-Entropy method. J Mech Sci Technol 23:2612–2623

Year of 2008

[1] An ZW, Huang HZ, Liu Y (2008) A discrete stress–strength interference model based on universal generating function. Reliab Eng Syst Saf 93(10):1485–1490
[2] Tian Z, Yam RC, Zuo MJ, Huang HZ (2008) Reliability bounds for multi-state *k*-out-of-*n* systems. IEEE Trans Reliab 57(1):53–58
[3] Tian Z, Zuo MJ, Huang H (2008) Reliability-redundancy allocation for multi-state series–parallel systems. IEEE Trans Reliab 57(2):303–310
[4] Huang HZ, Tao Y, Liu Y (2008) Multidisciplinary collaborative optimization using fuzzy satisfaction degree and fuzzy sufficiency degree model. Soft Computi 12, 995–1005
[5] Sun R, Huang HZ, Miao Q (2008) Improved information fusion approach based on DS evidence theory. J Mech Sci Technol 22:2417–2425

Year of 2007

[1] Huang HZ, Liu ZJ, Murthy DNP (2007) Optimal reliability, warranty and price for new products. IIE Trans 39(8):819–827
[2] Zuo MJ, Tian Z, Huang HZ (2007) An efficient method for reliability evaluation of multistate networks given all minimal path vectors. IIE Trans 39(8):811–817
[3] Liu ZJ, Chen W, Huang HZ, Yang B (2007) A diagnostics design decision model for products under warranty. Int J Prod Econ 109(1–2):230–240

Year of 2006

[1] Huang HZ, Zuo MJ, Sun ZQ (2006) Bayesian reliability analysis for fuzzy lifetime data. Fuzzy Sets Syst 157(12):1674–1686
[2] Huang HZ, Gu YK, Du X (2006) An interactive fuzzy multi-objective optimization method for engineering design. Eng Appl Artif Intell 19(5):451–460
[3] Fan XF, Huang HZ, Miao Q (2006) Evidence relationship matrix and its application to D-S evidence theory for information fusion. Lecture Notes Comput Sci 4224:1367–1373

Year of 2005

[1] Huang HZ, Tian Z, Zuo MJ (2005) Intelligent interactive multiobjective optimization method and its application to reliability optimization. IIE Trans 37(11):983–993
[2] Huang HZ, Wu WD, Liu CS (2005) A coordination method for fuzzy multi-objective optimization of system reliability. J Intell Fuzzy Syst 16(3):213–220
[3] Huang HZ, Li HB (2005) Perturbation finite element method of structural analysis under fuzzy environments. Eng Appl Artif Intell 18(1):83–91

Year of 2004

[1] Huang HZ, Tong X, Zuo MJ (2004) Posbist fault tree analysis of coherent systems. Reliab Eng Syst Saf 84(2):141–148