Proceedings of the 20th

AR² TS

Advances in Risk and Reliability Technology Symposium

21–23 May 2013

Edited by Lisa Jackson and John Andrews

Institution of MECHANICAL ENGINEERS

Safety and Reliability Society

Royal Statistical Society Founded 1834

Loughborough University

Proceedings of the 20th

Advances in Risk and Reliability Technology Symposium

21st – 23rd May 2013

Burleigh Court Conference Centre

Loughborough, Leicestershire, UK

Copyright © 2013

Published by:

Loughborough University
Loughborough
Leicestershire, LE11 3TU
United Kingdom

ISBN 978-1-907382 61 1

CONTENTS

Foreword 1
Lisa Jackson, John Andrews.

Programme 2

Keynote Presentations Abstracts 4

THE PAPERS

Predictive and Past Performance Assessment of Power System Reliability 6
Roy Billinton, University of Saskatchewan, Canada.

A System-wide Modelling Approach to Railway Infrastructure Asset Management 7
Dovile Rama, and John Andrews, University of Nottingham.

Analysis of the Contributions to the Performance of a Functional Product Design using Simulation 23
Sean Reed, Magnus Löfstrand, Lennart Karlsson,and John Andrews, University of Nottingham, and Luleå University of Technology, Sweden.

Modelling the Deferred Impact of Failures when Considering the Availability of Production Systems 43
Jelena Borisevic and Mark Rogers.

A Markov Modelling Approach to Railway Bridge Asset Management 55
Bryant Le, and John Andrews, University of Nottingham.

Fault Tree Analysis of Polymer Electrolyte Fuel Cells to Predict Degradation Phenomenon 75
Michael Whiteley, Lisa Bartlett-Jackson, and Sarah Dunnett, Loughborough University.

Maintenance and Planning in a Saudi Arabian Hospital 89
Hesham Alzaben, Chris McCollin, and Lai Eugene, Nottingham Trent University, and Riyadh Military Hospital.

Fault Diagnostics for Railway Point Machines 103
Marius Vileiniskis, Rasa Remenyte-Prescott, Dovile Rama, and John Andrews, University of Nottingham.

Probabilistic Analysis of Renewable Heat Technologies 120
Adam Thirkill and Paul Rowley, Loughborough University.

Quantifying Technical Risks: Insights into Theory-Practice Tensions in the Elicitation Process and Method 132
Gillian Anderson, Matthew Revie, and Lesley Walls, University of Strathclyde, Glasgow.

On Combined Data under Competing Risks 145
Tahani Coolen-Maturi, and Frank P. A. Coolen, Durham University, UK.

Towards a Failsafe Flight Envelope Protection: The Recovery Shield 160
J. A. Stoop, Lund University, Sweden.

The Art and Science of Whole Life Costing 172
Andy Kirwan and Julian Williams, Network Rail.

Localising Risk Estimates from the RSSB SRM 173
Chris Harrison, RSSB.

Use of a Generic Hazard List to Support the Development of Re-usable Safety Arguments in the Rail Industry 182
George Bearfield and Reuben McDonald, RSSB.

Automatic Construction of a Reliability Model for a Phased Mission System 192
K.S. Stockwell and S. J. Dunnett, Loughborough University.

Recent Advances in System Reliability using the Survival Signature 205
Frank P. A. Coolen, Tahani Coolen-Maturi, Abdullah H. Al-nefaiee, Ahmad M. Aboalkhair, Durham University, UK, and Mansoura University, Eygpt.

Degradation Test Analysis: A Case Study 218
Filippo De Carlo, Orlando Borgia, Mario Tucci, University of Florence, Italy.

A Petri-Net Modelling Approach to Rail Track Geometry Maintenance and Inspection 230
Matthew Audley, John Andrews, University of Nottingham.

Asset Management of a Railway Signalling System 244
Raphaëlle Barbier Saint Hilaire, Darren Prescott, John Andrews, University of Nottingham.

Modelling Railway Service Reliability 259
Claudia Fecarotti, John Andrews, Rasa Remenyte-Prescott, University of Nottingham.

Using Deep Belief Networks for Predicting Railway Operations Failures 274
Olga Fink, Ulrich Weidmann, Institute for Transport Planning and Systems, ETH Zurich, Switzerland.

Bayesian Analysis of Electric Transmission Network Outages 286
Tomas lešmantas, Robertas Alzbutas, Lithuanian Energy Institute, Kaunas.

Predictive and Diagnostic Analysis of a Holdup Tank by means of Dynamic Bayesian Networks 296
Daniele Codetta-Raiteri, Luigi Portinale, University of Piemonte Orientale, Alessandria, Italy.

Condition Monitoring Data in the Study of Offshore Wind Turbines' Risk of Failure 308
Maria C. Segovia,Matthew Revie, Francis Quail, University of Strathclyde, Glasgow.

Risk and Reliability: An Evolutionary Biologist's Perspective 320
Sara L. Goodacre, University of Nottingham.

Long-term Asset Maintenance Optimization at Scottish Water 321
Travis Poole, Tom Archibald, Jake Ansell, Robert Murray, Scottish Water Plc, Edinburgh.

Road Network Flow Modelling for Maintenance 330
Chao Yang, Rasa Remenyte-Prescott, John Andrews, University of Nottingham.

Probabilistic Reliability and Risk Analysis for Systems of Fusion Device 347
Roman Voronov, Robertas Alzbutas, Lithuanian Energy Institute, Kaunas, Lithuania.

Aleatory Uncertainty in Power System Reliability Index Assessment 356
R. Billinton, W. Wangdee, University of Saskatchewan, Canada, BC Hydro, British Columbia, Canada.

Choosing the Reliability Approach – A Guideline for Selecting the Appropriate Reliability Method in the Design Process 366
Cristina Johansson, Per Persson, Michael Derelöv, Johan Ölvander, Linköping University, Sweden, SAAB Aeronautics, Linköping, Sweden.

Investigating Electronics Reliability in Business Jet Applications 379
Ian James, Aero Engine Controls, Birmingham, UK.

The Dependability Case Is it Achievable 391
Richard Denning, Nick Barnett, Ministry of Defence Abbey Wood, Bristol, UK.

Onboard, Real-Time Detection of Adhesion Levels in the Rail/Wheel Contact 399
Peter Hubbard, Chris Ward, Roger Dixon, Roger Goodall, Loughborough University.

Use of Bayesian Updating to Combine Experts' Opinion and Results of Inspection in Bridge Management 411
Luis A. C. Neves, Dan M. Frangopol, University of Nottingham, Lehigh University, Bethlehem, Pennsylvania, U.S.

Stochastic State Space Methods for Railway Network Asset Management Modelling 421
Darren Prescott, John Andrews, University of Nottingham.

A Simple Model of the Software Failure Rate 434
Hendrik Schäbe

20th Advances in Risk and Reliability Technology Symposium (AR^{2}TS)

Foreword

A warm welcome to Loughborough and the Advances in Risk and Reliability Technology Symposium (AR^{2}TS). These proceedings represent the latest developments in the fields of risk and reliability, covering research as well as industrial case studies. The symposium aims to bring together like minded engineers and scientists striving to push forward the boundaries for implementation, development and improvement of risk and reliability techniques. It is hoped that the symposium will encourage creative discussion of traditional and innovative technologies, enable knowledge sharing and the formation of new collaborative opportunities.

This year's proceedings contain 34 papers, representing the work of authors across the industrial and academic domains. A broad spectrum of techniques are addressed in the areas of systems reliability assessment, hazard and risk analysis, maintenance planning and optimisation, fault diagnostics, data collection, asset management, software reliability, availability modelling and lifecycle costs. The discipline areas covered are power systems, railways, hospitals, road networks, electronics, water industry, design processes, jet engines, wind turbines, renewable heat, aircraft, and bridges.

There is an encouraging mix of academics and industrialists, from both the UK and overseas. This year sees authors from UK Universities in Durham, Edinburgh, Loughborough, Nottingham and Strathclyde. There are also authors from overseas Universities in Canada, Italy, Germany, Sweden, Netherlands and Switzerland. Companies who have contributed to research include Aero Engine Controls, BC Hydro (Canada), GL Noble Denton, the Lithuanian Energy Institute, , the Ministry of Defence, Network Rail, Riyadh Military Hospital (Saudi Arabia), RSSB, SAAB Aeronatics (Sweden), Scottish Water, and TÜV Rheinland InterTraffic GmbH (Germany).

We have the pleasure this year of three key note speakers. Professor Roy Billinton, from the University of Saskatchewan in Canada, will discuss the predictive and past performance assessment of power system reliability. Andy Kirwan, from Network Rail will discuss some of the latest approaches and issues in railway asset management. The final key note speaker will be Sara Goodacre, from the University of Nottingham, who will describe the mechanisms used by nature to enable survival and also touch on how biological means can be used to represent the failures occurring in engineering systems.

It is encouraging to see that young blood is taking on the challenge of research in this area, with a number of young researchers presenting their ideas this year. Prizes will be presented for the IMechE Best Young Researcher paper/poster and an award from the Safety and Reliability Society (SaRS) for the best practical paper.

Thanks go to the Programme Committee for their contributions to this symposium, the administrative support at Nottingham, and of course you the participants, whom we hope will contribute to a friendly and stimulating event. I hope you can join the Committee in the extra curricula activities of a wine reception (sponsored by SAGE) and conference dinner on the respective evenings.

Dr Lisa Jackson
Prof John Andrews

AR2TS Programme		
Tuesday 21st May		
10.00-10.10	**Introduction and Welcome**	John Andrews
10.10-11.00	**Keynote:** Roy Billinton (University of Saskatchewan, Canada) **Predictive and Past Performance Assessment of Power System Reliability**	Chair: John Andrews
11.00-11.15	**Coffee**	
11.15-12.45	Monte Carlo Methods	Chair: Frank Coolen
11.15-11.45	A System-wide Modelling Approach to Railway Infrastructure Asset Management	Dovile Rama and John Andrews
11.45-12.15	Analysis of the Contributions to the Performance of a Functional Product Design using Simulation	Sean Reed, Magnus Löfstrand, Lennart Karlsson, John Andrews
12.15-12.45	Modelling the Deferred Impact of Failures when Considering the Availability of Production Systems	Jelena Borisevic and Mark Rogers
12.45-13.45	**Lunch**	
13.45-15.00	**Poster Presentations**	Chair: Sarah Dunnett
13.45-14.00	A Markov Modelling Approach to Railway Bridge Asset Management	Bryant Le and John Andrews
14.00-14.15	Fault Tree Analysis of Polymer Electrolyte Fuel Cells to Predict Degradation Phenomenon	Mike Whiteley, Lisa Bartlett and Sarah Dunnett
14.15-14.30	Maintenance Planning in a Saudi Arabian Hospital	Hesham Alzaben, Chris McCollin, and Lai Eugene
14.30-14.45	Fault Diagnostics for Railway Point Machines	Marius Vileiniskis, Rasa Remenyte-Prescott, Dovile Rama, and John Andrews
14.45-15.00	Probabilistic Analysis of Renewable Heat Technologies	Adam Thirkill and Paul Rowley
15.00-15.30	**Poster Discussions and Coffee**	
15.30-17.00	**Risk and Safety Assessment**	Chair: John Catchpole
15.30-16.00	Quantifying Technical Risks: Insights into Theory-Practice Tensions in the Elicitation Process and Method	Gillian Anderson, Matthew Revie, Lesley Walls
16.00-16.30	On Combined Data Under Competing Risks	Tahani Coolen-Maturi and Frank P. A. Coolen
16.30-17.00	Towards a Failsafe Flight Envelope Protection: The Recovery Shield	John Stoop
6.00pm	Wine Reception (sponsored by SAGE)	
Wednesday 22nd May		
9.00-9.45	**Keynote:** Andy Kirwan, Julian Williams (Network Rail) **The Art and the Science of Whole Life Costing**	Chair: Lisa Jackson
9.45-11.15	**Rail Risk and Safety Assessment 1**	Chair: Lesley Walls
9.45-10.15	Localising Risk Estimates from the RSSB SRM	Chris Harrison
10.15-10.45	Use of a Generic Hazard List to Support the Development of Re-usable Safety Arguments in the Rail Industry	George Bearfield and Reuben McDonald
10.45-11.15	**Coffee**	
11.15-12.15	**Reliability Estimation 1**	Chair: Darren Prescott
11.15-11.45	Recent Advances in System Reliability using the Survival Signature	Frank Coolen, Tahani Coolen-Maturi, Abdullah H. Al-nefaiee, and Ahmad M. Aboalkhair
11.45-12.15	Degradation Test Analysis: A Case Study	Filippo De Carlo, Orlando Borgia, and Mario Tucci
12.15-13.15	**Lunch**	
13.15-14.30	**Poster Presentations**	Chair: Rasa Remenyte-Prescott
13.15-13.30	A Petri-Net Modelling Approach to Rail Track Geometry Maintenance and Inspection	Matthew Audley and John Andrews
13.30-13.45	Statistical Analysis to Reduce the Risk of Chest Injury for Older Occupants in Frontal Car Crashes	Karthikeyan Ekambaram, Richard Frampton, Lisa Bartlett
13.45-14.00	Asset Management of a Railway Signalling System	Raphaëlle Barbier Saint Hilaire, Darren Prescott and John Andrews

14.00-14.15	Modelling Railway Service Reliability	Claudia Fecarotti, Rasa-Remenyte-Prescott and John Andrews
14.15-14.30	Automatic Construction of a Reliability Model for a Phased Mission System	Kathryn Stockwell and Sarah Dunnett
14.30-15.00	**Poster Discussions and Coffee**	
15.00-17.00	**Analysis using Belief Network Approaches**	Chair: Jake Ansell
15.00-15.30	Using Deep Belief Networks for Predicting Railway Operations Failures	Olga Fink and Ulrich Weidmann
15.30-16.00	Bayesian Analysis of Electric Transmission Network Outages	Tomas Lešmantas and Robertas Alzbutas
16.00-16.30	Predictive and Diagnostic Analysis of a Holdup Tank by means of Dynamic Bayesian Networks	Daniele Codetta-Raiteri and Luigi Portinale
16.30-17.00	Condition Monitoring Data in the Study of Offshore Wind Turbines' Risk of Failure	Maria Segovia, Matthew Revie and Francis Quail
7.30pm	**Conference Dinner**	
Thursday 23rd May		
9.00-9.45	**Keynote:** Sara Goodacre (University of Nottingham) **Risk and Reliability: An Evolutionary Biologist's Perspective**	Chair: Lisa Jackson
9.45-10.45	**Asset Management**	Chair: Richard Denning
9.45-10.15	Long-term Asset Maintenance Optimization at Scottish Water	Travis Poole, Tom Archibald, Robert Murray and Jake Ansell
10.15-10.45	Road Network Flow Modelling for Maintenance	Chao Yang, Rasa Remenyte-Prescott and John Andrews
10.45-11.15	**Coffee**	
11.15-12.15	**Reliability Estimation 2**	Chair: John Andrews
11.15-11.45	Probabilistic Reliability and Risk Analysis for Systems of Fusion Device	Roman Voronov and Robertas Alzbutas
11.45-12.15	Aleatory Uncertainty in Power System Reliability Index Assessment	Roy Billinton and W Wangdee
12.15-13.15	**Lunch**	
13.15-14.45	**Reliability Case**	Chair: Lisa Jackson
13.15-13.45	Choosing the Reliability Approach – A Guideline for Selecting the Appropriate Reliability Method in the Design Process	Cristina Johansson, Per Persson, Michael Derelöv, and Johan Ölvander
13.45-14.15	Investigating Electronics Reliability in Business Jet Applications	Ian James
14.15-14.45	The Dependability Case is it Achievable	Richard Denning
14.45-15.15	**Coffee**	
15.15-16.15	**Railway Asset Management**	Chair: Luis Neves
15.15-15.45	Onboard, Real-Time Detection of Adhesion Levels in the Rail/Wheel Contact	Pete Hubbard, Chris Ward, Roger Dixon and Roger Goodall
15.45-16.15	Use of Bayesian Updating to Combine Experts' Opinion and Results of Inspection in Bridge Management	Luis Neves and Dan Frangopol
16.15-16.45	Stochastic State Space Methods for Railway Network Asset Management Modelling	Darren Prescott and John Andrews
16.45-17.00	**Awards and Close**	Richard Denning Lisa Jackson

Posters: 12 min presentation, 3 min change over
Presentations: 30 min slot - 25 min presentation, 5 min questions

Predictive and Past Performance Assessment of Power System Reliability

Roy Billinton, Power System Research Group
University of Saskatchewan, Canada

Electric power utilities collect considerable data on the past performance of individual system components and on how well the overall system performed it intended function. These data are also used to predict how the system will perform in the future and to evaluate the reliability of alternate expansion plans. This presentation will discuss a range of past reliability performance indices for bulk transmission and distribution systems and briefly illustrate the calculation of similar indices for predictive reliability assessment of future systems.

The Art and the Science of Whole Life Costing

Andy Kirwan and Julian Williams
Network Rail

Network Rail is one of the biggest asset management companies in the UK. In railway terms, we have the oldest system in the world and one of the busiest networks in Europe, with more train services than France, and more than Spain, Switzerland, The Netherlands, Portugal and Norway combined. We also have one of the safest rail networks, second only to Luxembourg in Europe, and one of the fastest growing, with a 50% increase in passenger journeys over the past decade and 30% more freight expected in the next five years.

The welcome increase in the demand for rail services presents a major challenge to asset managers – the people who plan and deliver work on the infrastructure. Additional trains increase the rate of asset degradation, restrict the time for access to the track to undertake preventive or restorative work, and exacerbate delays when failures occur. In parallel, there is a relentless drive for cost efficiencies, meaning the extra work needs to be done by fewer people.

To meet these challenges, we have had to rethink the way we prioritise work, to implement technological solutions that identify potential failures before they occur, and to devolve decisions to teams with a local understanding of the assets and close proximity to customers. To support this shift, we have introduced a whole life costing framework that puts customer service at the centre of decision making and provides consistency across asset disciplines and between business functions.

In this presentation, we will describe the approach taken, explain how it has been practically implemented, and show how the results have informed our investment plans for the next five years. Emphasis will be given to the models we have developed, the influence of uncertainties on decision making, and the compromises that are necessary to integrate ‘top down’ forecasts with ‘bottom up’ real world plans.

Risk and Reliability: An Evolutionary Biologist’s Perspective

Sara Goodacre
University of Nottingham

Evolutionary biologists study the process through which organisms adapt and survive. At the core of this process lies the generation of variation upon which natural selection acts. This can be viewed as an exploration of the different solutions that are possible for the same challenge (*i. e.* survival in a particular environment). The range

of solutions that is explored is rarely if ever exhaustive, being constrained by the time that an organism has had to adapt, and by the starting point, which is itself a product of previous evolutionary processes.

There are parallels between the evolutionary process described above and the search for optimum solutions in engineering designs. Survival (ie. non-failure of an engineered design) is maximised by searching for the optimal solution given known parameters. The search may or may not have been exhaustive and the 'solution' adopted is the best set of conditions found from those searched. There is a difference, however, in that evolution can and regularly does explore options of high risk whereas engineered solutions may not.

A study has been made of the literature on the evolution of bacterial and invertebrate genomes to ask to what extent the most successful solutions found by evolution to the challenge of survival favour redundancy, diversity or repair, or a combination of each of these.

Predictive and Past Performance Assessment of Power System Reliability

Roy Billinton, University of Saskatchewan, Canada

Electric power utilities collect considerable data on the past performance of individual system components and on how well the overall system performed it intended function. These data are also used to predict how the system will perform in the future and to evaluate the reliability of alternate expansion plans. This presentation will discuss a range of past reliability performance indices for bulk transmission and distribution systems and briefly illustrate the calculation of similar indices for predictive reliability assessment of future systems.

A System-wide Modelling Approach to Railway Infrastructure Asset Management

Dovile Rama and John Andrews
Nottingham Transportation Engineering Centre, University of Nottingham

Abstract

Increasing competition from alternative modes of transport and economic stimulus forces the railway industry to increase capacity and reduce expenditure. Without successful and effective management of infrastructure assets these goals would be difficult to achieve. Asset management is a decision making exercise which for complex infrastructure networks can be a challenging task. Various modelling techniques have been developed for investigation of alternative strategies to support the decisions making process.

In this paper a Petri Net methodology based approach is presented for modelling asset management of a railway line. The model presents degradation, inspection, maintenance and renewal of assets on the line. The approach has a hierarchical modular structure which allows representation and analysis of the rail infrastructure at different levels of granularity. The model is aimed to be used as a decision-aiding tool for railway asset management.

1. Introduction

An ever growing demand for more efficient and cost-effective railway infrastructure is driving an increasing interest in the application of asset management principles to the management of railway infrastructure [1]. Broadly speaking, asset management focuses on improving inspection, maintenance and renewal regimes in order to achieve required levels of infrastructure performance and dependability at the lowest costs. Managing these activities for aging and very diverse railway infrastructure is a challenging task. Various modelling techniques have been developed to support the process [2, 3, 4, 5].

In this paper a Petri Net (PN) based modelling framework is presented to be used as a decision-aiding tool for management of assets on the railway line. The main aim of this approach is to aid the coordination of activities and practices through which performance, risk and expenditure over the lifecycle of the railway line can be optimised. The modelling approach is described by using a track model as an example where asset inspection and maintenance regimes are modelled. Several studies on asset management of railway track have been performed. Some of them focused on particular activities which were a part of the whole asset management process such as maintenance [6] or renewal [7]. Patra et al [8] proposed a model that addressed different maintenance and renewal options, however their interdependencies were ignored. A number of techniques have been also developed to model the management process of individual elements of the track [9]. In order to

optimally utilise resources, minimise costs and achieve a desirable operating level, management of track monitoring activates, maintenance and renewal work as well as resources should be considered jointly. Furthermore, in order to effectively model the lifecycle of the track a system-wide approach needs to be considered. If track elements are considered individually activities involving all components, such as visual inspection of the track, cannot be taken into account appropriately. The model proposed in this paper has therefore been developed by taking a holistic approach and considers the track as a multi-component system.

Forecasting and planning of assets' future activities is a complex problem which involves both stochastic and deterministic processes. First of all, asset performance needs to be predicted taking into account the uncertain nature of their deterioration process [5]. Effectiveness of various intervention actions, such as inspections or repairs, also often has a degree of uncertainty. Activities performed periodically over the lifetime of assets, such as regular inspections or preventive maintenance, are usually represented as a deterministic processes. Since Petri Nets are capable of modelling both, stochastic and deterministic process, as well as their interactions, this techniques has therefore been chosen to be used for developing an asset management model for a railway line.

2. Track

The track which is a fundamental part of the railway infrastructure can be considered as a multi-component system. Standard railway track has two rails supported on sleepers. Rails and sleepers are often referred to as track superstructure. The sleepers themselves are laid on crushed stone ballast.

Rails are manufactured using steel of various grades. Discrete pieces of rail can be joined by either bolted fish plates or welded to form longer track sections. Degradation mechanisms that limit the serviceable and operational life of rails can be categorised into Rolling Contact Fatigue (RCF) and rail breaks and defectives including loss of rail profile. Examination of the rails to detect wear and locate defects is performed by visual inspection and ultrasonic testing. On discovery of any rail degradation evidence maintenance actions undertaken may involve hand-grinding, weld repair or replacing the effected piece of rail. Train-born grinding performed on a cyclic basis is widely recognised as the principal preventive measure to control wear and slow the growth of rail defects [10].

Sleepers are used as a base for the track. There are four different types of sleepers: hardwood, softwood, concrete and steel sleepers. They provide the following functions [11, 12]:

- Load transfer from the rails to the ballast;
- Securing the rails to maintain constant rail spacing;
- Mechanical strength in resisting both the lateral and the longitudinal movement of the rails.

Railway sleepers may develop various defects over time. They may rot, brake or split and therefore become less capable of meeting performance requirements. Sleepers which are unable to support the rails or retain track gauge are classed as ineffective. Damaged and deteriorated sleepers are being replaced with new ones.

Rails are secured on sleepers using fastenings. There is a big variety of types of fastenings used on the railway; including spikes and screws, rail supports (chairs, plates) etc. Over time the fastenings can also become defective, loose or go missing. Measures taken in order to return track condition to a satisfactory state may include tightening of the fastenings or replacing them along with the affected sleepers. Sleepers and fastening requiring intervention are normally detected by a means of visual inspections.

The ballast, which is a formation of loose and coarse grained aggregate, supports the superstructure of track by resisting its vertical, lateral and longitudinal displacement and dissipating forces transmitted by the sleepers. Furthermore, it governs the level of the rails and provides water drainage from the track structure. With time ballast deficiencies, excess or voiding can develop over small lengths of track. Under traffic loading, or as a result of certain maintenance interventions, ballast can break down into small aggregate with an accumulation of fines. This results in poor track geometry and uneven support of the railway [13]. Minor ballast defects in small sections will usually be treated by regulating, packing and clearing ballast, and clogged and polluted ballast will be dug out and re-ballasted [14]. Track geometry in a longer track section is usually improved by the means of tamping and stoneblowing. Routine visual inspection is used to reveal signs of ballast deterioration. Regular automated measurement of the track geometry is also a part of the inspection routine [15].

3. Petri Nets

Petri Nets (PN) [16] are a powerful modelling formalism combining graphic representation with a mathematical background used to represent complex behaviour of dynamic systems. They have a very broad spectrum of applications and have also been used for developing asset management models [17, 18].

A PN is a graph containing two types of nodes called places and transitions. Places, drawn as circles, are often used to represent a particular state of the system or activity being modelled. Transitions, drawn as bars, are events or actions which represent the dynamic behaviour of the system such as change of its state or performance of a certain activity. Places are linked to transitions and vice verse by directed edges or arcs. There can be multiple arcs of the same direction between the place and the associate transition and vice verse.

Objects called tokens representing the current system state move between places through transitions according to transition switching or firing rules. A transition becomes enabled for switching when all its input places are marked, i.e. the number of tokens each place contains is at least equal to the arc

multiplicity going from the place to the transition. In the standard PN when switching happens, either after a prescribed deterministic or random delay time, 1) each input place loses a number of tokens determined by the associated arc multiplicities and 2) each output place receives a number of tokens determined by the associated arc multiplicities.

A special kind of arc, the inhibitor arc, is used to disable switching of the transition. The presence of a token in the input place with the inhibitor arc will prevent the associated transition from firing as shown in Figure 1.

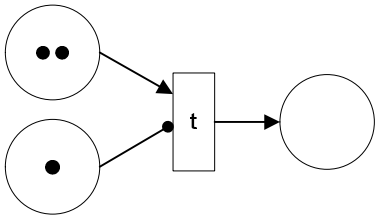


Figure 1. PN with an inhibit arc.

A basic PN can be further extended by adding additional features in order to improve the efficiency of the PN when adopting the technique to solve a specific problem. For example, a number of non-standard transition types have been used to build PN models in [9, 19], namely a place condition transition, a reset transition, a decision making transition and a convolution transition. These transition types have also been utilised in the model presented in this paper by enhancing some of their originally presented functionality. Specifically, a delay time greater or equal to zero has been implemented in the switching rule of a reset transition. A delay time before firing the place conditional transition can now be either deterministic or random which is then sampled from a specified distribution.

Additionally, a new type of transition has been introduced which was named a place priority transition. This type of the transition is used when several transitions share the same input place(s) and priority to fire one of the activated conflicting transitions needs to be determined. The decision as to which transition to fire is based on specified rules and is dependent upon the number of tokens in associated input places which are referred to as conditional input places. A multi-functional transition type has also been introduced in the current model. A transition of this type can combine features of some of the earlier named non-standard PN transitions.

4. Modelling Asset Management Decisions and Activities

4.1 Modelling Approach Overview

There is a great diversity among individual railway lines in terms of assets utilised. Due to variable environmental conditions and railway traffic loads which impact asset degradation processes, the condition of the infrastructure among individual lines may also be very different. Recognising that individual lines and their parts can deteriorate and therefore be managed in a number of

ways, the framework was developed for modelling the state of the line and its different asset management options using a hierarchical modular approach.

The top level, railway line module, is comprised of a number of section modules. The section module models a single 1/8th mile section of the railway. It is further broken down into smaller modules each one representing a single asset. Using the railway track as a simplified representation of the railway line, the modules of the three elements of the track, namely rails, sleepers with fastenings and ballast, will form a section module. Each component module has several sub-modules which are used to model asset condition degradation, inspection regimes and intervention options, accounting for availability of personnel and machinery resources.

The model introduced is developed using the PN formulation. Through utilisation of PN subnets commonly shared among modules of the same and higher levels of modularity the hierarchical concept of the modelling architecture is maintained. The conceptual framework of the modelling approach is presented in Figure 2.

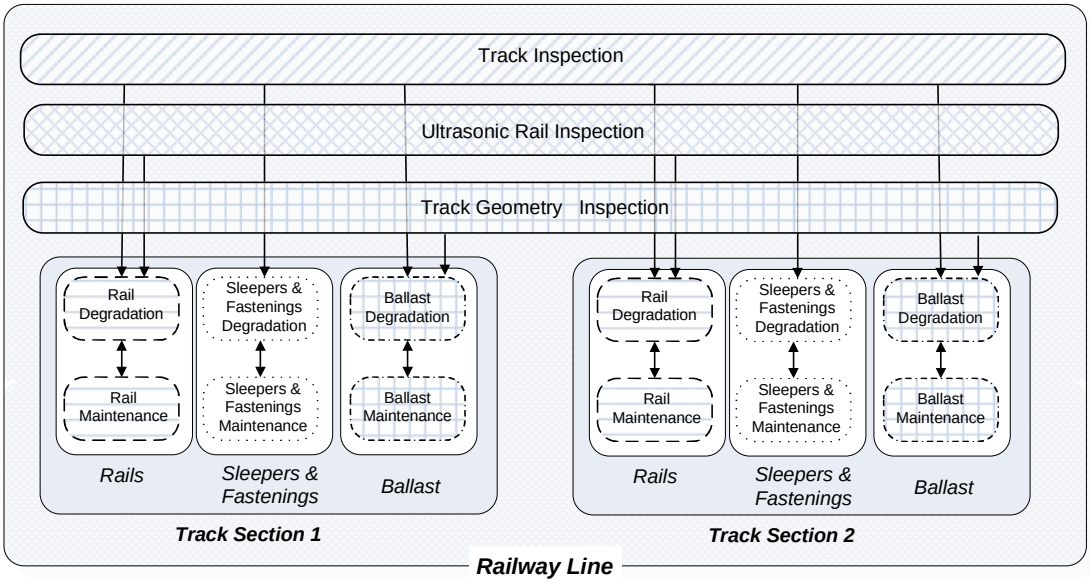


Figure 2. Track Lifecycle Modelling Framework

4.2 Asset Petri Net Modules

At the first layer of modularity in the modelling architecture proposed, the entire railway line is divided into asset PN modules. These modules represent individual types of assets utilised on a 1/8th mile railway section. Here a PN module for sleepers and fastenings will be presented in detail. PN modules of the ballast and rails are built in a similar manner and therefore will not be discussed in detail. Attention will only be drawn to instances where significant differences among asset modules exist.

In further discussion of the PN modules the places and transitions will be labelled for the purpose of identification and short descriptions will be

provided for places to clarify the state of the system they represent. Places and transitions are labelled in a particular manner. Each label comprises of two numbers. The first number represents one of the following sub-modules: 1 - asset degradation/failure sub-module, 2 – inspection regime sub-module, 3 – sub-module for manual maintenance activities which do not require special machinery, 4 – sub-module for maintenance performance using machinery (not included in the sleepers and fastening module), 5 – sub-module for management of personnel resources. The second number determines the level of modularity, i.e. places numbered from 1 to 199 relate solely to activities within a track section. Places numbered from 200 to 299 represent activities within a single line and places numbered from 300 are used to model activities at a region level that consists of a number of connected railway lines. Additionally two more numbers can be added to form a label of the node in order to distinguish nodes that belong to a particular section and/or line modules.

Degradation and Inspection Sub-modules: A PN structure representing the degradation process of sleepers and fastenings is presented in Figure 3. Places numbered from P1_1 to P1_4 represent possible condition states of the track sleepers and fastenings in a $1/8^{th}$ mile section. Rather than considering the condition of individual assets the overall condition of sleepers and fastenings in the section is modelled which is determined by the urgency of intervention. For example, if any sleepers or fastenings develop a defect place P1_2 will be marked indicating that repairs in this section is of low urgency. If, however, any of sleepers or fastenings become ineffective maintenance in this section becomes more urgent and place P1_3 is marked. The delay time for the stochastic transitions between the places is derived from a random sample taken from the appropriate distribution. The distributions can be obtained from historical data and further processed if necessary [19].

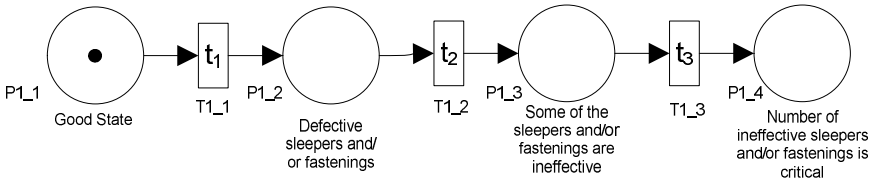


Figure 3. Degradation sub-module

The inspection regime is modelled using the inspection sub-module as shown in Figure 4. When inspection is underway place P2_1 is marked. Then depending on the marking of places P1_2 - P1_4 one of the transitions T1_4 - T1_6 is enabled. The enabled transition fires immediately and reveals the current state of sleepers and fastenings. After a time period ε_{2_1} (the time period over which the inspection takes places) transition T2_2 fires marking place P2_2 and indicating the end of the inspection. The next inspection begins after a time period θ_{2_2} marking place P2_1 and the process is repeated. If at that time maintenance in underway in the section (place P3_10 is marked), transition T2_1 will be inhibited and inspection will be postponed.

Additional inspections are carried for rails and ballast. Ballast geometry is checked regularly using a measurement train and ultrasonic testing is performed to detect rail defects. Inspection sub-modules for these assets will therefore have an additional loop added, similar to that shown in Figure 4.

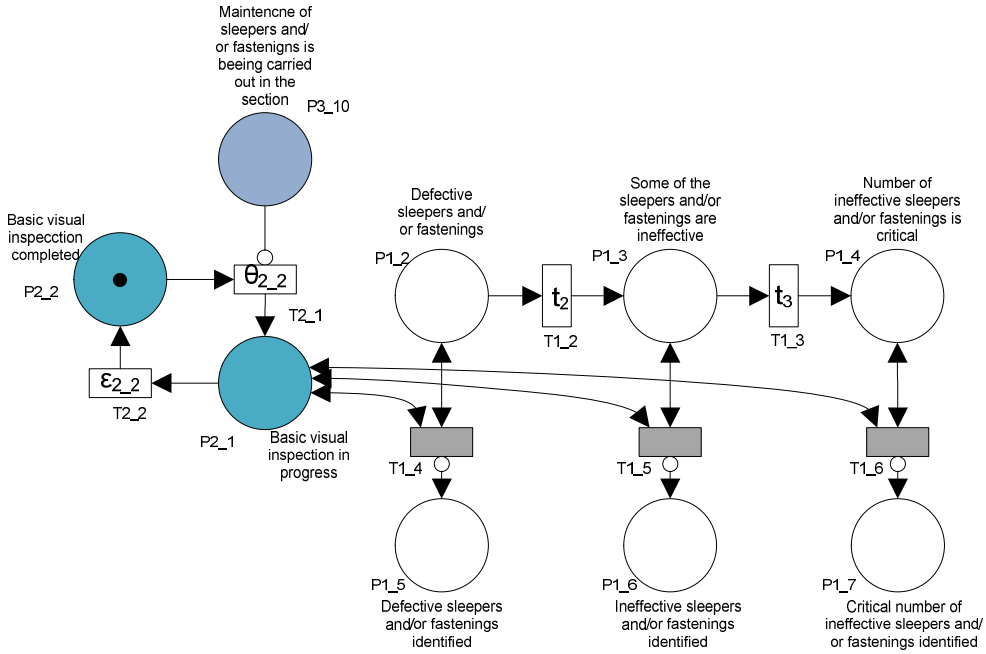


Figure 4. Inspection of sleepers and fastenings

Defective sleepers or fastenings found need to be replaced with new ones. It is therefore assumed that any intervention will return the state of section sleepers and fastening into a good state. The PN fragment for modelling the process is presented in Figure 5.

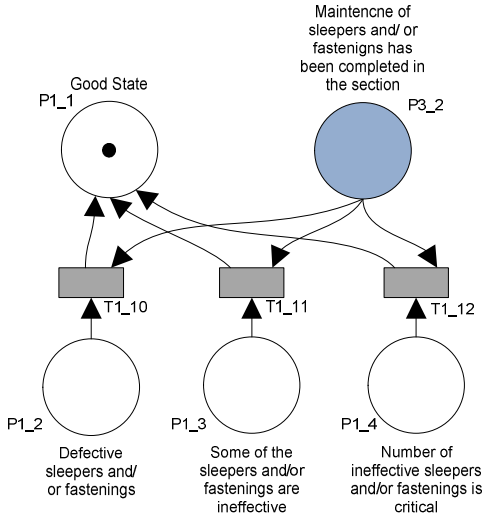


Figure 5. Updating current state of section sleepers and fastenings

Maintenance Sub-module: Figure 6 depicts the PN structure modelling the repairs of ineffective sleepers and/or fastenings whose number have not yet

reached a critical level. Once defective sleepers and/or fastenings have been found, the timescales for repairs are determined. This is modelled by placing a single token in places P3_10, P3_3(s) and P5_4. Here labelling (s) is used to identify that the place is commonly used in the PN of other assets. After the planned delay (denoted by θ_{3_3}) when the maintenance is due to be performed (place P3_12 is marked) a request is placed for personnel resources (a single token is placed in P5_12). The request is of a medium priority which is identified by placing 4 additional tokens in place P5_4. The lists of maintenance activities are also updated by placing a single token in places P3_201, P3_203 and P3_5(s) simultaneously. Once the request has been fulfilled a token occurs in place P5_16 and transition T3_15 is enabled and fires immediately. Places P3_1 and P3_14 become marked signifying that repair work is being carried out in the section. Transition T3_15 is a decision making transition and by firing it the demand for maintenance actions in the current section and the whole line is updated. Work is completed after a time period equal to ε_{3_5} and place P3_15 becomes marked. If sleepers and/or fastenings can be repaired earlier than scheduled, as part of the concurrent or opportunistic maintenance regimes (which are not presented in Figure 6), place P3_13 would be marked. This subsequently would enable transition T3_14. Resulting changes in the marking of the PN would occur in the same manner as those after firing transition T3_15.

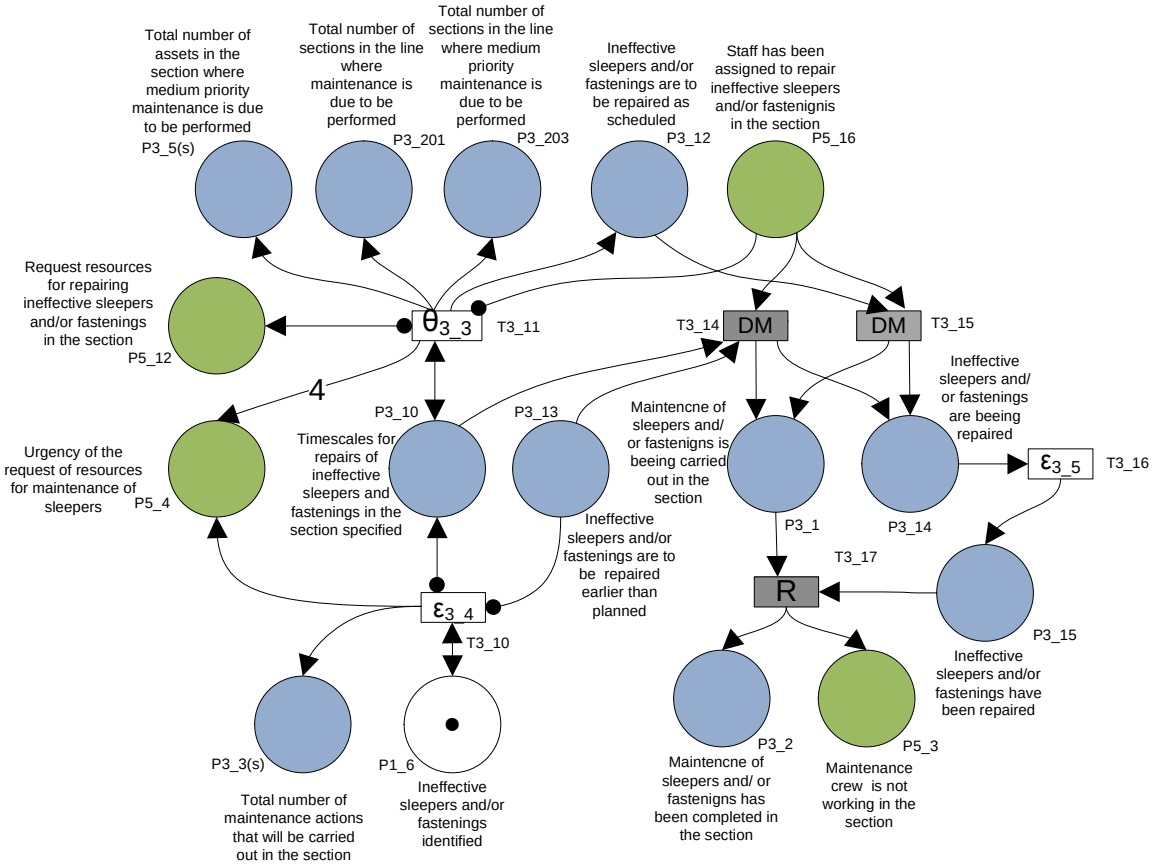


Figure 6. Maintenance of ineffective sleepers and fastenings

Similar PN structures exist for modelling the maintenance activities once defective sleepers and/or fastenings have been found and when the number of ineffective assets reaches a critical level.

Resource Allocation Sub-module: Two illustrative PN structures for modelling the allocation of staff resource to carry out specific tasks in the section are presented in Figure 7. Management of the resources to repair minor defects of sleepers and/or fastenings as part of regularly planned maintenance activities is implemented in Figure 7a. A maintenance team which has been made available to carry out the maintenance of sleepers (place P5_7 is marked) will be assigned to carry out defect repairs once requests for resources to repair defects (place P5_11 marked) and to carry out routine maintenance (place P5_10 marked) have been placed. In this case transition T5_5 becomes inhibited and transition T5_6 fires marking places P5_14 and P5_15. If maintenance staff have only been requested for routine maintenance then only transition T5_5 is enabled and fires immediately marking place P5_14. Transitions T5_5 and T5_6 are decision making transitions and on firing update the state of urgency of personnel requests for maintenance activities.

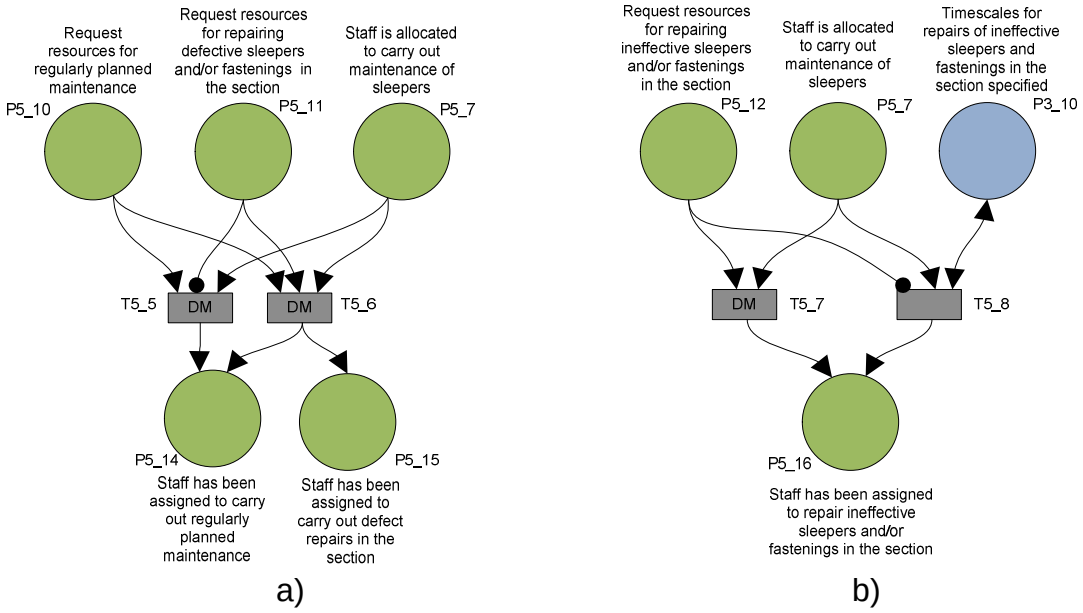


Figure 7. Allocation of staff resources for specific tasks

As shown in Figure 7b, ineffective sleepers and/or fastenings will be repaired at the time requested (identified by placing a token in place P5_12) once maintenance team becomes available to work in the section (place P5_7 is marked). In this case transition T5_8 is inhibited and enabled transition T5_7 fires marking place P5_16. This is a decision making transition with an additional rule introduced to change the marking of place P5_4 by removing 4 tokens. If for any reason a maintenance team has been made available to work in the section earlier than planned, i.e. places P5_7 and P3_10 are marked but place P5_12 is unmarked, work will then be carried out earlier than planned and transition T5_8 will be enabled and fire. A similar PN

structure exists for allocation of staff resources to repair a critical number of ineffective sleepers and/or fastenings.

Since machinery, i.e. stoneblowers and tampers, can also be utilised for ballast maintenance additional PN structures exist to model the availability and coordination of these resources.

4.3 Building a Section Module

As mentioned earlier, a single section module is built by joining the asset modules together. The section module, however, cannot be viewed simply as a collection of discreet modules assembled together. A module integration strategy is needed in order to address the asset management principles at a system level and to take into account existing interdependencies among asset management activities.

First possible interactions among individual asset sub-modulus are addressed that will occur when a section module is constructed. We assume that the degradation of each component is independent of the state of other components. Thus in the section module each asset will have an individual PN degradation sub-module. All track components are inspected during the basic visual inspection and therefore a common visual inspection PN sub-module will be linked to each component degradation sub-module as discussed earlier. Additionally, a ballast geometry inspection sub-module will only be linked to a ballast degradation sub-module and so will an ultrasonic rail inspection sub-module which will be linked to a rail degradation sub-module.

With regards to maintenance, while any repairs or renewals are carried out according to component-specific policies, the timing of interventions among different assets can be coordinated in order to optimise the maintenance activates. For this purpose additional nodes are added in the asset maintenance sub-modules. Additional places are introduced in each maintenance sub-module to indicate which other assets can be maintained concurrently. As shown in Figure 8, in the sleepers and fastenings sub-module the added places are labelled as P3_41 and P3_81. The marking of these places signifies that ballast and rail repairs can be carried out while ineffective sleepers and/of fastenings are being maintained. Equally, maintenance of sleepers and fastenings can also be rescheduled. Figure 8 demonstrates a PN subnet that models the maintenance of ineffective sleepers and fastenings initiated as part of the concurrent maintenance activities (place P3_11 is marked). Here enabled transition T3_11 is a place conditional transition whose firing delay time depends on the marking of place P3_11. When P3_11 is marked transition T3_11 fires immediately and places P3_12 and P5_16 become marked. Subsequently, the marking of these places enables transition T3_12 which when fires places a token in place P3_13. The changes in the marking of the PN that follow are the same as those discussed in the description of the PN subnet presented in Figure 6.

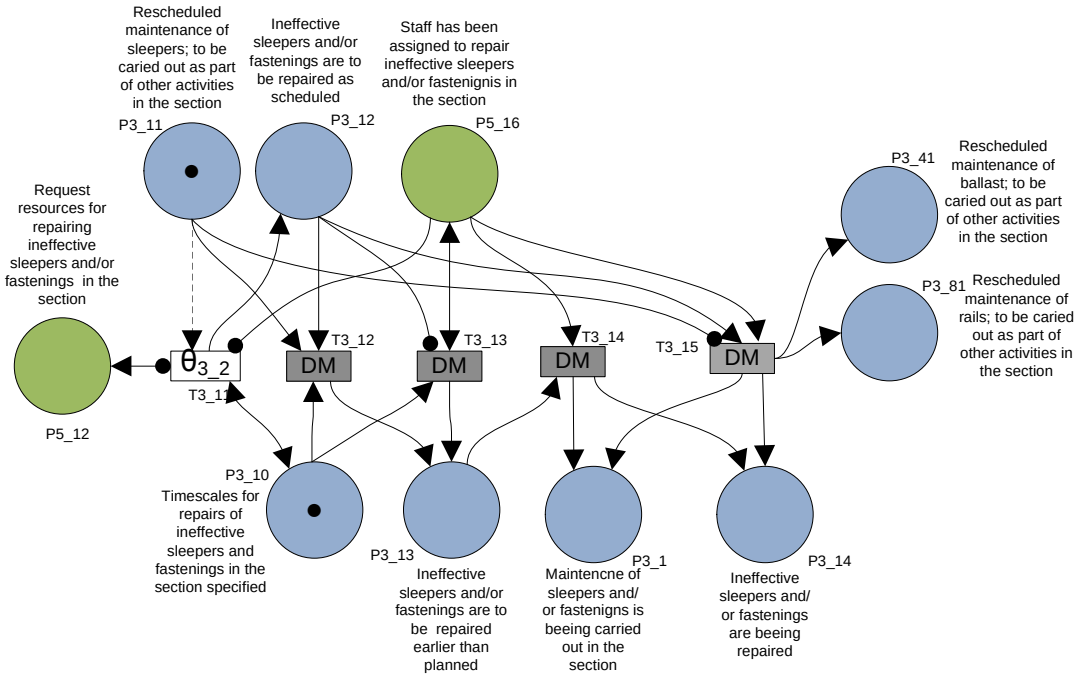


Figure 8. Coordination of maintenance activities in the section

Since resource allocation is directly linked to the maintenance demand, the assignment of specific intervention tasks also needs to be coordinated, especially when there is several assets whose maintenance is due to be performed. For this purpose an additional PN structure is utilised as shown in Figure 9 Transitions T5_2 – T5_4 are place priority transitions. It means that the order of firing these transitions, when there is more than one transition enabled, is determined based on marking of places P5_4 – P5_6. That is a transition whose conditional input place (identified with a dashed arc) has the most tokens will fire first.

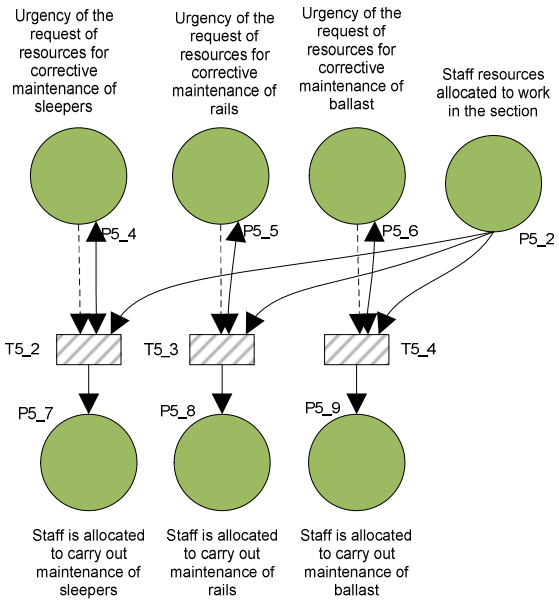


Figure 9. Allocation of staff resources for maintenance of a specific asset type

4.4 Building a Railway Line Module

Similarly to a section module, a line PN module is constructed by joining individual section PN modules. It involves integrating as many copies of a single section PN module as there are sections in the line in question and several additional PN subnets. Due to existing interdependencies among assets and their operation only some of asset sub-modules will appear in every section module, while others will have only a single copy and will be shared among all section modules. For example, if it is assumed that the track components in a section degrade independently from components in other sections, each section module will contain as many degradation sub-modules as there are assets in the section. Considering that the inspection activities of asset conditions are carried out simultaneously in all sections of the line, only single copies of three inspection sub-modules will be included and will be linked to asset degradation sub-modules in each section PN module. Integration of the maintenance modules is achieved through the introduction of common places in each section module. Some of them, such as P3_201 and P3_203 were shown in PN structures in Figure 6. These places represent lists of maintenance activities that are due to be performed and their level of urgency at a line-level.

The management of maintenance activities within a line, especially the ones that have the same time schedules, is achieved through coordinated allocation of staff resources. When staff resources are limited decisions on the prioritisation of specific tasks in particular sections are made taking into account the overall condition of the line as well as conditions of individual assets within each section. Some sections in the line maybe more critical than others and this also has to be taken into account. PN structures used for this purpose are presented in Figures 10 and 11.

A PN in Figure 10 models a decision making process for allocate of available staff resources on either of two sections on the line. The decision is made using two place priority transitions T5_1_1 and T5_1_2. Priorities for firing the transitions are determined based on the marking of associated conditional places. The rules can be determined so that to implement specific intervention strategies. For example, the number of assets which are due to be repaired and whose maintenance is of the same level of urgency can be compared among sections. Alternatively, the highest number of maintenance actions that are due to be performed can determine the priority for maintenance among the sections.

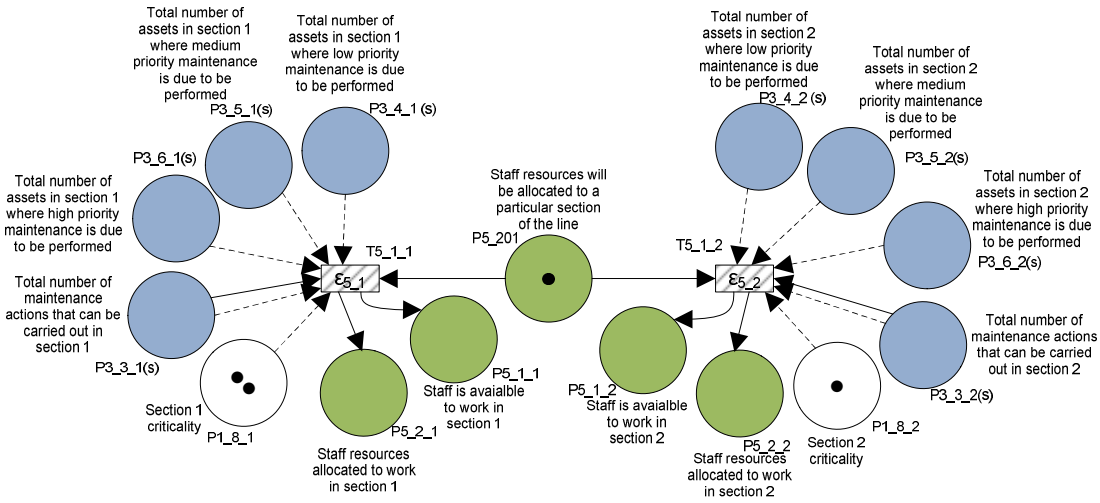


Figure 10. Identifying priority of maintenance among sections

Figure 11 shows a PN structure which models how a decision is made whether staff remains working in the same section or whether they have to start work in another section of the line. Transition T5_11 is a decision transition which fires once staff have completed the task assigned, i.e. when places P5_3_1 and P5_1_1 are marked and place P5_2_1 is unmarked. Depending on the marking of conditional input places identified with dashed arcs, the marking of places P5_1_1, P5_2_1 and P5_201 will change. The assumed strategy for the decision making is as follows. If the urgency of the maintenance request in the current section is not lower than that in the remaining sections of the line then a token will be placed in places P5_1_1 and P5_2_1 signifying that the maintenance team will continue working in the same section. If, however, there are sections with a higher urgency of maintenance, then place P5_201 will be marked and a decision will have to be made considering the overall state of the line as shown in Figure 11. These rules can be easily modified to implement alternatives strategies for allocation of resources.

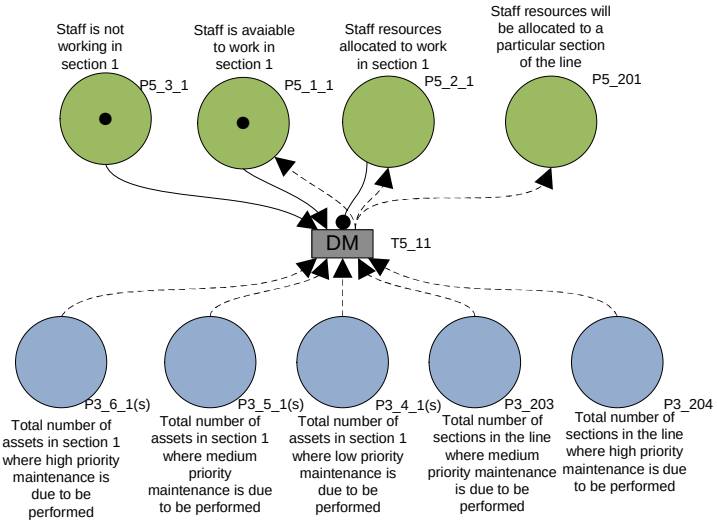


Figure 11. Identifying location of following maintenance activities.

4.5 Possible Model Enhancements

The modular architecture described provides a great flexibility allowing the construction of PN models representing different parts of the railway line. Several extensions can be made to the model presented. A higher level of modularity, i.e. region module, can be achieved by linking line modules and by following similar integration rules at the ones provided. This module would provide capability of investigating the impact of various asset management decisions in a much larger area of the network by taking into account existing variability among smaller parts of the network.

The current example discussed considered only track assets on a railway line. A great diversity of other types of assets are utilised on the line, including signalling, power provision, telecommunication equipment, structures and so on. Having a library of PN modules of these individual assets and by choosing the asset modules relevant to individual sections of the network a very detailed and realistic infrastructure model can be developed.

5. Conclusions

A PN based approach has been presented for modelling an operation lifecycle of railway assets, specifically asset degradation, inspection and maintenance and renewal strategies. The latter strategies are modelled taking into account the asset condition, resources availability and industry policy requirements. The PN methodology has been chosen due to its capability to model complex processes allowing for the integration of different elements of the lifecycle in a more realistic fashion. New types of PN transitions have been proposed both to simplify the PN structures and to enable the modelling of complex intervention strategies which otherwise would be difficult and in some cases impossible to model using conventional PN structures.

The approach is structured in a hierarchical modular fashion, from individual assets forming a particular railway asset group or system in a $1/8^{th}$ mile railway section to railway lines represented as a series of connected sections, and potentially even larger regions of the railway network. Such a structure allows the modelling of the operation and management of railway infrastructure taking into account existing variation among individual network parts. Models can be tailored to take into account infrastructure features, intervention strategies and resource availability specific to particular parts of the network. It provides a great flexibility in modelling a railway line of any size e.g. an individual section or the whole route, and in any part of the network taking a system-wide approach. Furthermore, it enables the optimisation of intervention strategies by providing concurrent and opportunistic maintenance options.

The models then can be used to investigate the effects of asset management strategies in terms of costs and infrastructure dependability. For this purpose a Monte-Carlo simulation procedure would be performed to obtain the

relevant metrics such as numbers of asset failures, train delay minutes, maintenance intervention demand for maintenance recourses and so on.

In the paper the modelling approach was demonstrated using railway track as the only set of assets on the railway line. The model can be easily extended by introducing other assets, .e.g. signals, points etc., in order to develop a more comprehensive asset management model of a railway line.

Acknowledgement:

John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is also Director of The Lloyd's Register Foundation (LRF)[1] Centre for Risk and Reliability Engineering at the University of Nottingham. Dovile Rama is the Network Rail Research Fellow in Asset management. They gratefully acknowledge the support of these organizations.

References

1. Guidelines for the Application of Asset Management in Railway Infrastructure Organisations, In Kirwin, A. and Gradinariu, T. (eds), Rail System and Communications Department Report. Paris: International Union of Railways (UIC), September 2010.
2. Carretero, J., et. al., Applying RCM in large scale systems: a case study with railway networks, *Reliability Engineering & System Safety*, 82(3): 257-273 (2003).
3. Márquez, F.P.G., et. al.,Life Cycle Costs for Railway Condition Monitoring, *Transportation Research Part E: Logistics and Transportation Review*, 44(6): 1175-1187 (2008).
4. Camci, F., Comparison of Genetic and Binary Particle Swarm Optimization Algorithms on System Maintenance Scheduling Using Prognostics Information, *Engineering Optimization*, 41(2):119-139 (2009)
5. Márquez, A.C; The maintenance management framework: models and methods for complex systems maintenance. Springer: 2007, p.333.
6. N. Rhayma, N., et. al. A probabilistic approach for estimating the behaviour of railway tracks, *Engineering Structures*, 33: 2120-2133 (2011)
7. Andersson, M., et. al. Estimating the marginal cost of railway track renewals using corner solution models, *Transportation Research Part A*, 46: 954-964 (2012)
8. Patra, A. P., Söderholm, P., Kumar, U., Uncertainty in Life Cycle Cost of Railway Track, *Proceedings of Reliability and Maintainability Symposium, 2008. RAMS 2008: 42-47*.
9. Prescott D., Andrews, J., A railway Track Ballast Maintenance and Inspection Model for Multiple Track Sections. *Proceedings of PSAM 11 & ESREL 2012*.

[1] The Lloyd's Register Foundation (LRF) supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

10. Network Rail Standard – *Inspection and Maintenance of Permanent Way*, NR/L2/TRK/001/mod10, Issue 6, December 2012.
11. Profillidis V.A., Railway Management and Engineering, 3rd ed,. Ashgate Publishin Limited, pp 473 (2006).
12. Manalo A. et al, A review of alternative materials for replacing existing timber sleepers, *Composite Structures*, 92: 603-611 (2010).
13. Lam, H.F. and Wong, M.T., Railway Ballast Diagnose through Impact Hammer Test, *Procedia Engineering*, 14: 185-194 (2011).
14. Historical Network Rail Standard – *Inspection and maintenance of permanent way – Installation requirements, maintenance limits and intervention limits*, NR/L2/TRK/001/E01, Issue 4, December 2009.
15. Audley, M. and Andrews, J.D., The Effects of Tamping on Railway Track Geometry Degradation, *submitted to Proc. Of the IMechE, Part F: J. Rail and Rapid Transit*.
16. Murata, T., Petri Nets: Properties, Analysis and Applications, *Proceedings of the IEEE*, 77 (4): 541-580 (1989).
17. Fouathia, O., et al. Stochastic approach using Petri nets for maintenance optimization in Belgian power systems, Proceedings of 8th International Conference on Probabilistic Methods Applied to Power Systems, 2004: 168-173.
18. D'Addio, G.F., Savio, S. and Firpo, P., Optimized Reliability Centered Maintenance of Vehicles Electrical Drives for High Speed Railway Applications, Proceedings of ISIE'97: 555-560.
19. Andrews, J.D., A Modelling Approach to Railway Track Ballast Asset Management, *Proc. of the IMechE, Part F: J. Rail and Rapid Transit*. 227(1): 56-73 (2012).

Analysis of the contributions to the performance of a functional product design using simulation

Sean Reed[1], Magnus Löfstrand[2], Lennart Karlsson[2], John Andrews[1]
NTEC, University of Nottingham, Nottingham, UK
The Faste Laboratory, Luleå University of Technology, Luleå, Sweden

Abstract

Functional products (FP) consist of combined hardware, software and support services that are sold to the customer under performance-based contracts that guarantee a specified level of functional availability. The supplier is responsible for the development, manufacture, support and upgrade of a FP during the contract period. In comparison to a traditional hardware sale only contract, an FP transfers risk from uncertain availability and support costs from the customer to the supplier. This is a major advantage for the customer but means that the supplier must understand and optimise the availability and support costs of a FP design. During product development, simulation can be used to analyse potential FP designs, predict how they will perform and identify possible areas for improvement – providing vital qualitative and quantitative decision support. In this paper, a methodology for analysing a FP design to predict how it will perform and determine the contribution of individual elements of the FP to its overall performance is described. This methodology is then applied to analyse an example of a FP.

1. Introduction

Functional products [1] consist of an integrated package of hardware and support services. They are sold under performance based contracts where the supplier retains ownership of the hardware and the supplier's compensation is tied to the value that the product generates for the customer [2]. An example from the private sector is the 'power-by-the-hour' scheme offered by Rolls-Royce PLC for the supply of gas turbine engines to airlines. One of the key factors influencing the value generated for the customer is the availability of the critical functions it provides, where availability is defined as the proportion of uptime [3]. Developing a product that delivers its critical functions at high availability and predictable cost, preferably as low as possible, is therefore essential for a functional product supplier. The provision of maintenance is central to achieving this since it plays a key role in both the attainment of desired functional availability levels and in the total costs of delivering the function.

The development of a functional product design is generally an iterative process that involves the conception and evaluation of numerous designs and design choices for both hardware and support services [4]. It is important that the design is improved as much as possible during product development since modifications made during the implementation stage are, in general, much more costly and difficult to implement.

This paper investigates the application of a functional product modelling tool, which has been developed by the authors, to the guiding of hardware and maintenance process improvement activities during the critical product development stage. It begins with a description of the methodology developed by the authors for predicting FP performance through simulation and the introduction of some analysis techniques for determining the contribution of various elements of the FP design to its performance. The application of the methodology to an example FP is then demonstrated.

2. Methodology

In this section, a methodology based on simulation modelling for analysing the efficiency of a FP support system, identifying the waste and ranking the maintenance tasks according to their contribution to the FP performance is described. An overview of the methodology is shown in

Figure 1.

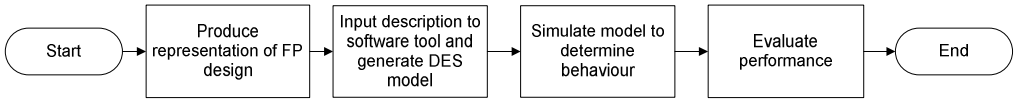


Figure 1 – Overview of the methodology.

2.1. Representation of a Functional Product Design and Conceptual Model

The first part of the methodology is a conceptual model for representing a FP design. In order to determine the performance of an FP in terms of availability and support costs, there are four main elements that must be represented in the conceptual model: the hardware reliability, the maintenance processes, the maintenance strategy and the maintenance resource logistics.

Hardware: There are two levels to the conceptual model of the hardware reliability. At the first level are the models of the individual component reliabilities and at the second level are the models of the system reliabilities. The reliability model for a component is represented by a Petri net [5] featuring a place for each state of the component (i.e.. working or the presence of failure modes) and transitions associated with distributions that correspond to the transition times between states. The Petri net must be designed such that a token is in a single component state place at any time, representing its current state, and an additional place labelled ‘repair’ where a token can be inserted to trigger an immediate transition to the working state (i.e. model the restoration of the component to the as-new condition). Example Petri nets for various types of component reliability model are shown in

Figure 2, where filled transitions are instantaneous and transition times for non-filled transitions are associated with a distribution.

The reliability of a sub-system or system is a function of the reliabilities of the components from which it is comprised, with failures occurring when certain

component failure combinations occur. The fault tree technique [3] is used to represent the reliability structure of the system, describing the logical relationship between component failure events and higher level failures.

Maintenance Processes: Maintenance procedures can be defined as sequences of tasks or actions that apply resources (labour, spare parts, tools and facilities) to hardware items (systems, sub-systems and components) for the purpose of extending their lifetime (time to failure) or restoring them to the working (often as-new) condition. They can be represented as a graph structure where the nodes represent the tasks and the edges represent the sequencing constraints. The task nodes with outgoing edges that connect to a certain task node represent the prerequisite tasks for the latter. A task is initiated as soon as any set of prerequisite tasks with incoming edges connecting to the task node at the same point, named a prerequisite set, are completed. Each node can also have multiple outgoing edges, each representing a different possible outcome. These outcomes can be associated with a probability of occurrence (in which case they are mutually-exclusive) or depend on the state of a component or presence of an event in a fault tree. The former is useful for representing the possibility of errors during maintenance, whilst the latter is useful for representing the outcome of diagnostic tasks. This graph structure has been named as a Maintenance Procedure (MP) Graph [6][7]. Each maintenance task is associated with a distribution that models its completion time and the resources it utilises, consumes and produces.

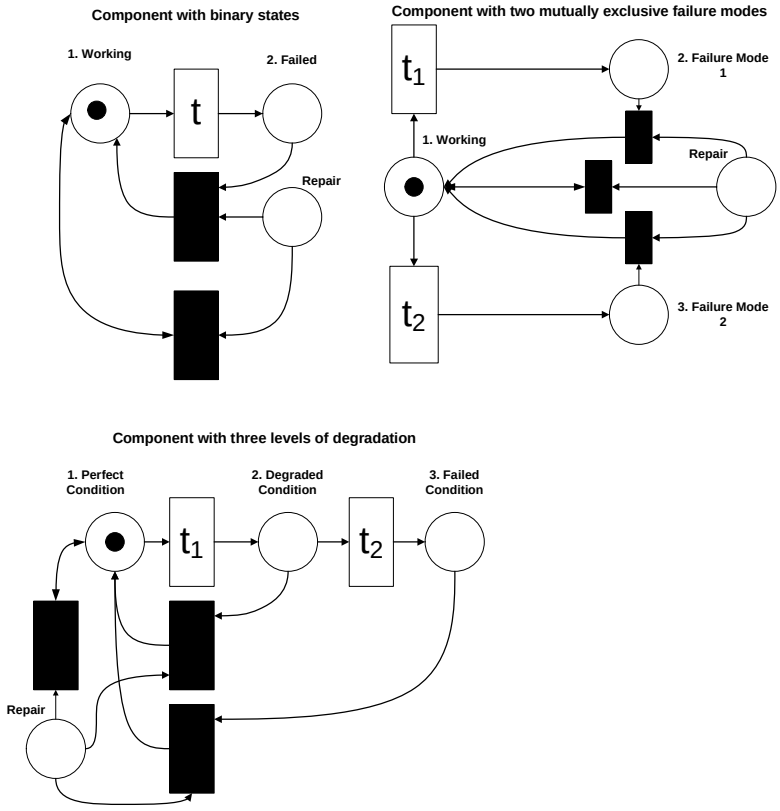


Figure 2 – Petri net representations of various component reliability models.

Maintenance Strategy: The maintenance strategy is represented as a set of maintenance strategy elements, where each element consists of a maintenance procedure and a trigger condition. The trigger defines the condition under which the maintenance procedure is initiated and may be an event from a fault tree (corrective or reactive maintenance), a component age or time since last initiated (preventive maintenance) or the condition of a hardware item (predictive maintenance).

Maintenance Resource Logistics: Forming a conceptual representation of the maintenance resource logistics, which determines the delay in the arrival of the resources required by the maintenance tasks within procedures, for a general FP is difficult due to the wide variety of differences in the logistical system that may be employed by an FP supplier in practice. Often the assumption that the logistical delay for obtaining a particular resource follows a certain distribution that is independent of demand is suitable, but in certain cases a more complex modelling approach is required.

2.2. Computer Model Generation and Simulation

The second step in the methodology is to analyse the behaviour of the conceptual model so that predictions on its performance, such as functional availability and support costs, can be obtained. For this purpose, the authors have developed a software tool in the C# programming language that generates a discrete event simulation [8] representation of the conceptual model from a set of input files describing the elements of the FP design outlined in the previous section. This model enables the probabilistic behaviour of the FP Design to be generated over a specified period of operation (e.g. 5 years). Each such replication constitutes a simulation trial and a large number of trials can be performed to produce a statistical picture of its behaviour. During each trial, any model variables of interest (such as component functional state variables or maintenance task activity variables) can be monitored and the state changes and times recorded within a database. Performing a greater number of trials enables more sample data to be obtained and hence greater confidence in the accuracy of the statistics calculated during the model output data analysis. However, it also increases the computational expense of the simulation and therefore the number of simulation trials carried out is a compromise between these two factors.

2.3. Model Output Data Analysis

Once the computer model has been simulated and data collected on its operational behaviour, the final step in the methodology is to apply post-simulation processing to calculate statistics and metrics relating to the FP design performance. These statistics can then be used to determine the performance level of the FP design and to provide decision support that aids the improvement of that design. A huge range of metrics can be calculated for a FP design by the software tool developed by the authors, from high level metrics that indicate overall performance to low level metrics which can help identify design problems and possible improvements. Some of the metrics that can be calculated are shown in

Figure 3. In this section, the functional availability and total support costs are defined as these are the metrics which a FP provider must optimise for. Some lower

level analysis methods that quantitatively measure the contribution of individual elements of the FP design to the FP performance are introduced. Since improving areas of the design that contribute most to the FP performance will often result in the greatest performance gains, these methods can provide valuable decision support for design development.

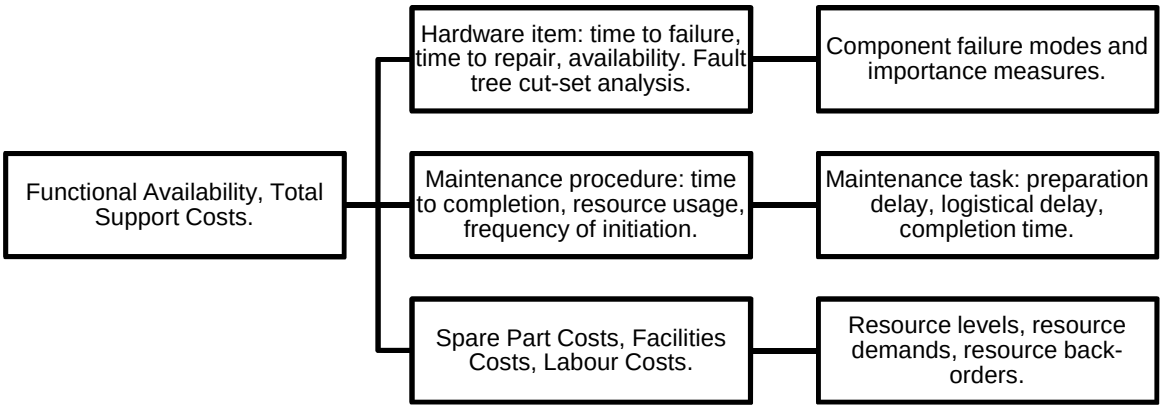


Figure 3 – Hierarchy showing some of the high and low level metrics that can be calculated from the data obtained through simulation of the FP model.

Availability and Total Support Costs: The availability and total support costs are the metrics that a FP supplier optimises for. Whilst they are not helpful for decision support in improving a particular design, they show the overall performance of a design.

The functional availability for the ith FP, A_i, within a particular time period is defined as its functional uptime during that period, U_i, divided by the sum of the uptime and downtime, D_i, as shown in Eqn. (1).

$$A_i = \frac{U_i}{U_i + D_i} \tag{1}$$

The cost of compensating the customer of the ith FP for failure to achieve the guaranteed availability, C_{A_i}, is given by Eqn. (2) where G is the guaranteed availability, *T* is the duration of the FP support contract and *F* is the compensation paid per unit downtime above the guaranteed maximum.

$$C_{A_i} = \begin{matrix} 0 & if\ A_i \geq G \\ (G - A_i) \times T \times F & otherwise \end{matrix} \tag{2}$$

The total costs of supporting a FP, from a reliability and maintenance perspective, is the sum of the compensation paid to customers for failure to achieve guaranteed functional availability levels and the costs of providing maintenance. Eqn.

(4) gives the total costs, C_T, of supporting *N* instances of a FP where C_{S_i} are the costs associated with providing maintenance for the ith FP.

$$C_T = \sum_{i=1}^{N} C_{A_i} + C_{S_i} \tag{3}$$

Measuring the Efficiency of the Support System: The support system within a FP applies maintenance processes to the FP hardware to increase its availability. Analysis and improvement of the support system performance is therefore vital during the development of an FP. Various process improvement concepts have been developed within different fields. Methods for improving production processes were originally developed within the Toyota Production System [9] at Toyota in Japan and later formed the basis for a group of techniques known as lean manufacturing [10]. These techniques were later adapted by Harrington [11] to the streamlining of business processes under the names of business process improvement, redesign and reengineering. Common to these and other process improvement methodologies is the requirement to understand the processes (both qualitatively and quantitatively), identify and eliminate waste (including wasted time, labour and material) and prioritise improvements based on the potential for adding value. A measure of the efficiency of the support system is defined here as the proportion of total maintenance hours (where a maintenance hour is defined as an hour spent by a maintenance engineer in the application of a maintenance procedure) utilised within the support system that are spent performing value-added tasks, where a value-added task is defined as a task that directly results in the restoration or extension in the expected lifetime of a hardware item. This efficiency measure, *E*, is given in Eqn.

(4) where *V* is the total maintenance hours spent performing value-added tasks and *T* is the total maintenance hours spent within the support system on maintenance.

$$E = \frac{V}{T} \tag{4}$$

The remaining time within the support system, i.e. the non-value added time, can be further categorised as follows:

- Travel time – time spent transporting personnel, spare parts, tools and equipment during maintenance.
- Re-work – time spent fixing defects or correcting errors made during maintenance.
- Diagnostics – time spent determining the causes of failure and appropriate remedial actions.

- Setup time – time spent preparing for the performance of the value-added tasks or returning the system to its operational condition afterward.
- Documentation/Bureaucracy – time spent filling in forms, reports etc.
- Waiting – time spent waiting to perform a task (idle time).

Analysing the proportion of time spent in these different areas can help identify the types of FP design changes that could be made to reduce non-value added work and thus improve the efficiency of the support system.

Contribution of Component Reliability and Maintenance Task Time to FP Performance: Some metrics for measuring the contribution of component reliability and maintenance task completion time to the FP performance will now be introduced. In order to use these metrics an appropriate indicator of the FP performance must be chosen first, i.e. the objective function for which the supplier is aiming to optimise. For a FP this will:

- Be a function of the product performance in terms of functional availability and support costs.
- Account for the variability in performance, which should be minimised to reduce exposure to the downside risk, as well as the expected performance.

Reed et al [12] suggested the use of a metric based on the predicted return on investment and risk aversion of the supplier as an appropriate objective function.

Component importance measures [13] are widely used to measure the contribution of the reliability of a component to system reliability performance. The contribution of component failure to the objective function, r_i, can be calculated for component *i* through Eqn.

(5), where O_f is the objective function to be minimised and $F_i(t)$ is the cumulative probability of failure for component *i* at time *t.*

$$r_i = 1 - \frac{O_f | F_i(\infty) = 0}{O_f} \tag{5}$$

Similar to the metric given, the contribution from maintenance task completion time to FP performance can be defined. The contribution of the task completion time to the objective function, b_i, can then be calculated for maintenance task *i* through Eqn.

(6), where O_f is the objective function to be minimised and $G_i(t)$ is the cumulative probability that task *i* is completed at time *t* after initiation.

$$b_i = 1 - \frac{O_f | G_i(0) = 1}{O_f} \tag{6}$$

Components and maintenance tasks with the highest values according to these measures have the greatest contribution, or importance, to the FP performance and may be considered as strong candidates for improvement. However, the

contributions of individual elements of the FP design often depend on design decisions elsewhere in the design and this should be taken into consideration. For example, the contribution of a component to the system performance according to the definition given in Eqn.

(5) will depend not only on its reliability but also the speed with which failures are restored through maintenance.

3. Application to an Example Functional Product

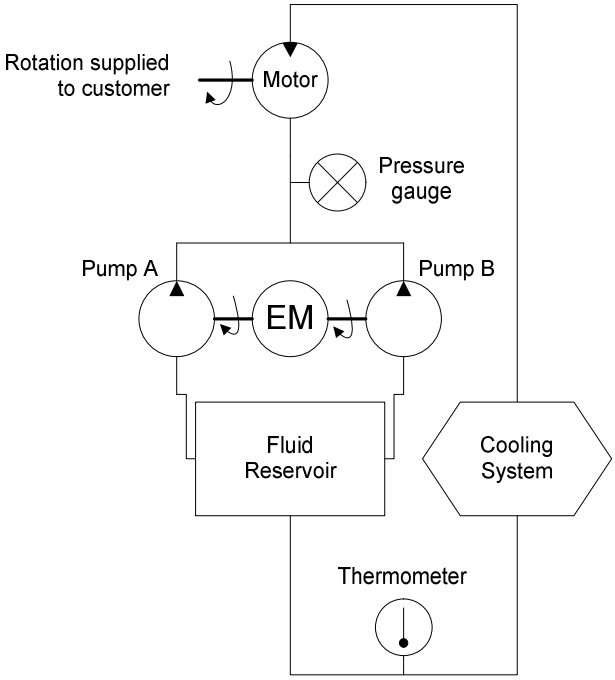


Figure 4 – Schematic of hydraulic drive system.

An example FP design that will be used to demonstrate the application of the methodology that was described in Section 2 is introduced in this section. All units relating to time are given in hours. The example system is a simplified closed-loop hydraulic drive system that provides rotational torque and speed. As shown by the schematic in

Figure 4, it comprises of an electric motor, two hydraulic pumps (named pump A and pump B), a hydraulic motor, a cooling system, a fluid reservoir, a pressure gauge and a thermometer. The electric motor powers the pumps which pressurise the hydraulic fluid which is then converted back into mechanical rotation by the hydraulic motor. The hydraulic motor is the same component as the two hydraulic pumps but mounted in the reverse configuration. The functionality of the fluid pressurisation system (comprising of the electric motor and hydraulic pumps) is monitored by a pressure gauge that measures the pressure of the hydraulic fluid within the supply line to the motor. The cooling system removes heat from the hydraulic fluid to prevent overheating which can damage the system components such as seals. Its functionality is monitored by the thermometer which measures fluid temperature within the supply line to the fluid reservoir. The components in the system have a

single failure mode with a failure time represented by a Weibull distribution with the parameters shown in

Table 1, with the exception of the pressure gauge, thermometer and fluid reservoir which are all assumed to be perfectly reliable.

Component	Scale Parameter	Shape Parameter
Hydraulic Pump/Motor	25000	2
Electric Motor	40000	1
Cooling System	40000	1.3

Table 1 – Component failure time Weibull distribution parameters.

The fault tree for the top event of “Loss of Rotation” is shown in

Figure 5. Note that the system features some degree of redundancy since a single pump functioning is sufficient to provide the required pressurisation of the hydraulic fluid.

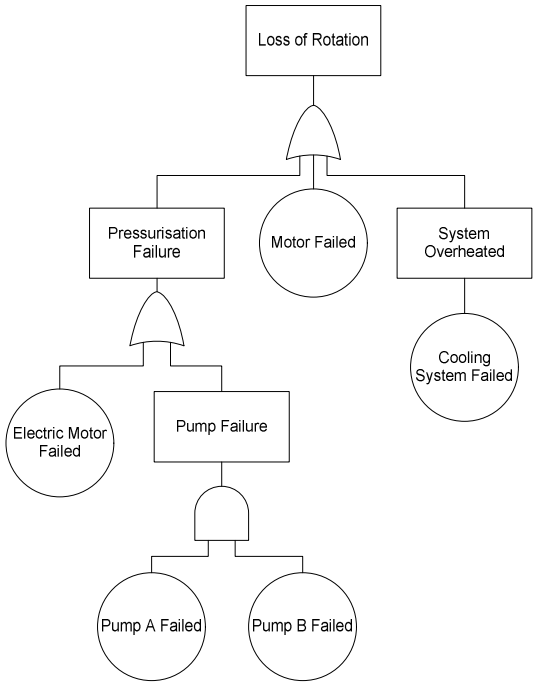


Figure 5 – Fault tree for “Loss of Rotation” top event for the example hydraulic drive system.

The designs for the maintenance processes are shown in

Figure 6 and

Figure 7.

Figure 6 shows the process for the restoration of the pressurisation system,

Figure 7A shows the process for the restoration of the hydraulic motor and

Figure 7B shows the process for the restoration of the cooling system. The maintenance tasks within the processes are numbered on the MP graphs and are

assumed to have completion times that are distributed according to the triangular distribution with the minimum, modal and maximum times for each as shown in

Table 2.

Table 2 also gives the number of maintenance engineers utilised by each maintenance task and a classification for the task type according to the following key: A=Value-added, B=Setup, C=Diagnosis, D=Re-Work, E=Documentation. In this paper, the logistical delays in obtaining certain maintenance resources are included explicitly within the maintenance procedure representations.

Task number	Description	Minimum time	Mode time	Maximum time	Number of Engineers	Cost	Task Type
1	Maintenance engineers travel between SS base and install site.	1.00	1.50	2.00	2	1	F
2	Engineers sign-in at reception	0.10	0.25	0.50	2	0	E
3	Engineers travel between reception and system location	0.10	0.15	0.20	2	0	F
4	Shut-off power supply	0.05	0.10	0.15	1	0	B
5	Open motor cabinet	0.05	0.08	0.10	1	0	B
6	Attach test equipment to motor	0.20	0.25	0.30	1	0	B
7	Diagnose motor failure	0.05	0.08	0.10	1	0	C
8	Disconnect test equipment	0.20	0.25	0.30	1	0	B
9	Fetch replacement motor	0.20	0.30	0.50	1	0	F
10	Replace electric motor	0.50	0.75	1.00	2	50	A
11	Close motor cabinet	0.05	0.08	0.10	1	0	B
12	Re-connect power supply	0.05	0.10	0.15	1	0	B
13	Update maintenance report	0.05	0.08	0.10	1	0	E
14	Remove pump maintenance panel and pump	0.20	0.25	0.30	2	0	B
15	Diagnose pump failure	0.10	0.15	0.20	1	0	C
16	Replace pump	0.15	0.20	0.25	2	0	B
17	Fetch replacement pump	0.20	0.30	0.50	1	0	F
18	Prepare and fit replacement pump	0.30	0.35	0.40	2	30	A
19	Test pump function	0.10	0.15	0.20	1	0	C
20	Correct installation error	0.20	0.25	0.30	2	0	D
21	Replace maintenance panel	0.05	0.08	0.10	1	0	B
22	Engineers sign and hand-in maintenance report	0.10	0.25	0.50	2	0	E
23	Drain hydraulic fluid	0.20	0.25	0.30	1	0	B
24	Disconnect cooling system	0.40	0.50	0.70	2	0	B
25	Fetch replacement cooling system	0.20	0.30	0.50	1	0	F
26	Replace cooling system	2.00	2.50	3.00	2	20	A
27	Re-fill hydraulic fluid	0.25	0.30	0.35	1	1	B

Table 2 – Maintenance task details.

The following modelling assumptions are made:

- There is sufficient availability of all other maintenance resources required to perform the tasks, i.e. the tools, facilities and spare parts.
- The system is shut-down when the system top event of functional failure occurs and deterioration of any working components is paused until functionality is restored.

Table 3 describes the maintenance strategy for the example hydraulic drive FP.

Maintenance Procedure	Trigger
See Figure 6.	"Pressurisation Failure" event in fault tree (see Figure 5).
See Figure 7A.	Hydraulic Motor component fails.
See Figure 7B.	"System Overheated" event in fault tree (see Figure 5).

Table 3 – Maintenance strategy for the example hydraulic drive FP.

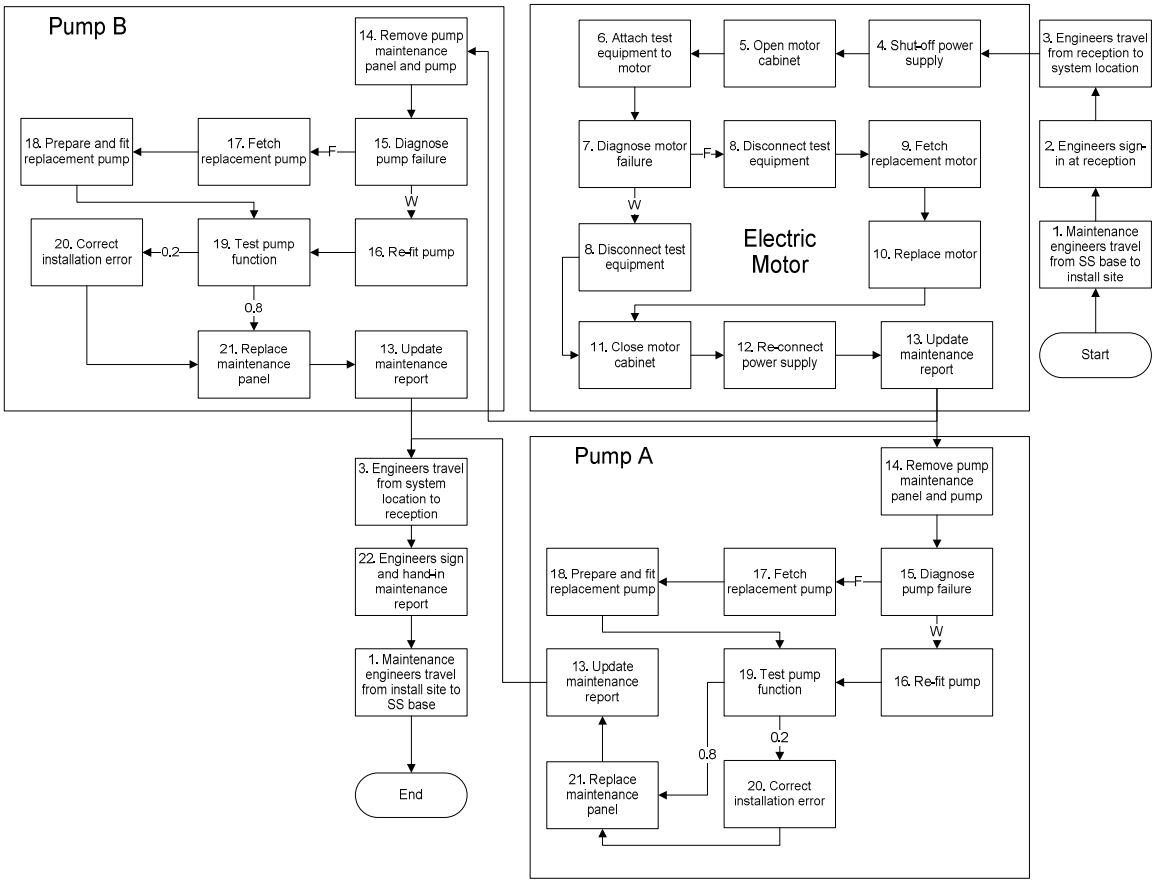


Figure 6 – Maintenance procedure for the diagnosis and restoration of the electric motor and hydraulic pumps within the example hydraulic drive FP.

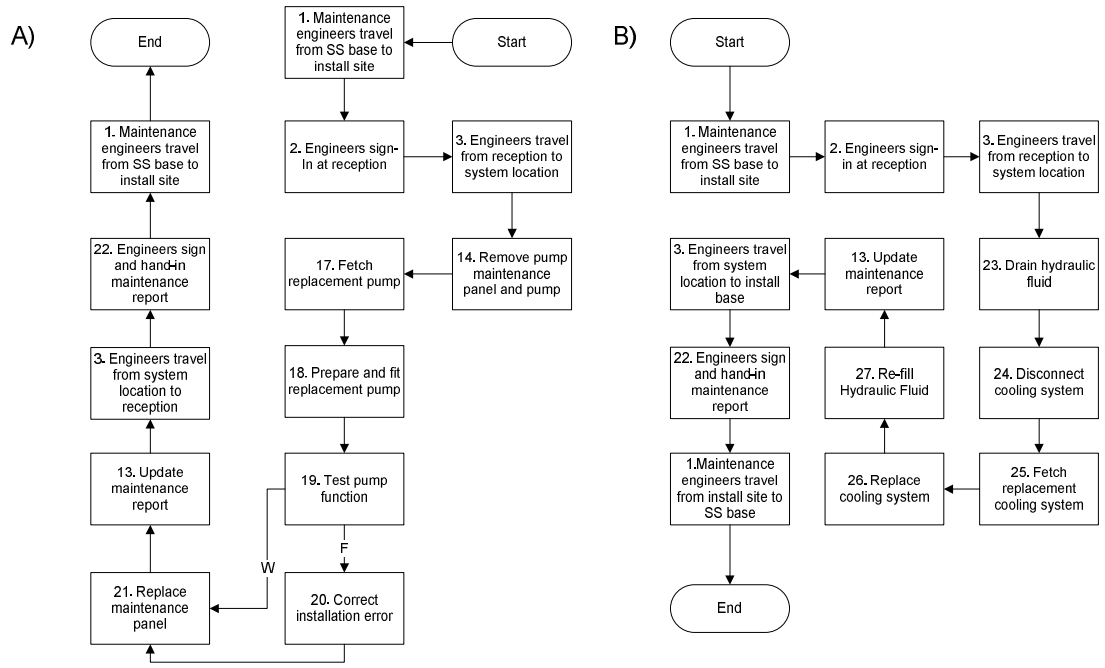


Figure 7a (left) and Figure 7b (right) – Maintenance procedures for the restoration of the (a) hydraulic motor and (b) cooling system within the example hydraulic drive FP.

For the analysis of this FP, the following assumptions are also made:

- Twenty FPs are sold to customers at time 0 and supported over a period of five years (43,800 hours).
- Each customer is guaranteed that each FP will achieve an availability of 99.9% or better over the 5 year period, i.e. a maximum total of 43.8 hours of downtime. The distribution of the downtime over the period is not important.
- For every additional hour, or part thereof, of downtime above the guaranteed maximum, compensation of 100 will be paid by the customer to the supplier.
- There is a cost of 1 unit per hour that a maintenance engineer is performing maintenance.

The direct costs of performing each task (not including maintenance engineer's time) are given in

- Table 2.
- The objective considered by the supplier is to minimise the expected total costs from supporting the FP and compensation for failures of the FP to achieve the guaranteed availability over the support period.

4. Results and Discussion

The results of applying the methodology described in Section 2 to the example FP described in Section 3 are presented in this section.

Figure 8 shows a plot of the cumulative probability for total compensation costs for the example FP. Since the compensation costs at the 50th percentile are just under 2000, equating to a mean additional downtime above the guaranteed maximum of 100 hours for each of the twenty supported FP, the supplier might consider reducing the guaranteed availability or increasing the availability through design improvements in order to reduce these costs.

Figure 9 shows a plot of the cumulative probability for total maintenance support costs for the example FP and

Figure 10 shows a plot of the cumulative probability for total support costs for the example FP. As shown, maintenance support costs of between 5000 and 5500 can be expected but may be as high as 6000, whilst total support costs of between 4500 and 8500 are typical. There is also a probability of approximately 20% that total support costs will be greater than 8500 and a small chance that costs will be 11000 or above. The FP provider may therefore wish to improve the FP design to reduce the variability in the potential cost to reduce the risk of an adverse outcome.

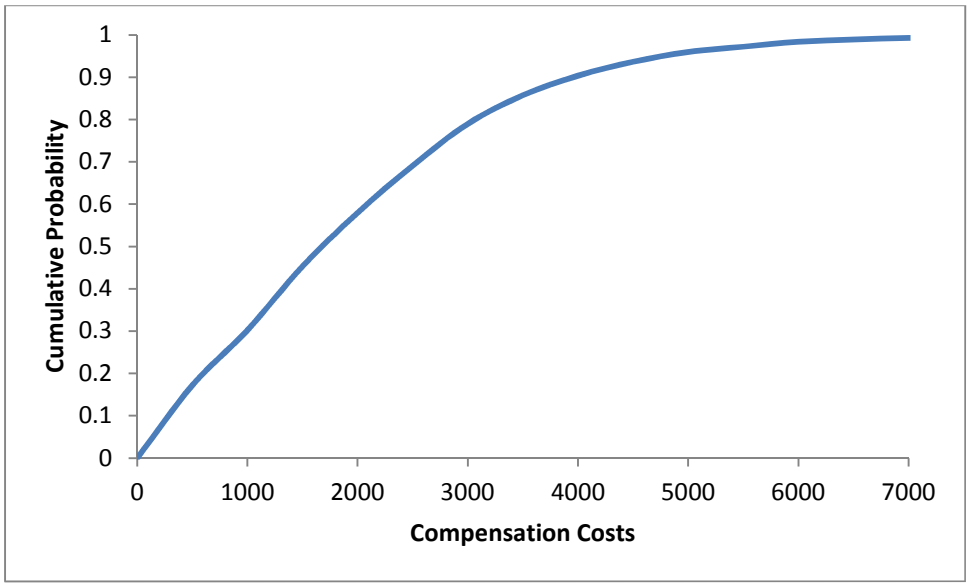


Figure 8 – Plot of the cumulative probability for total compensation costs for the example FP.

The task types in which maintenance engineer hours are expended in the support of the FP are shown by the histogram in

Figure 11. This shows that almost 50% of the total engineer hours are spent travelling to and from the maintenance location or fetching spare parts, over 10% of time is spent waiting and just under 10% of time on setup tasks. The high transport percentage suggests that the maintenance support could be made more efficient by locating engineers on-site or closer to the customer locations and storing spare parts closer to the FP installation locations. The high waiting percentage suggests that a re-design of the processes to include greater concurrency in tasks or better balance workloads would also benefit the efficiency of the FP support provision. The amount of time spent on setup tasks could be reduced by redesigning the hardware to

increase maintainability, for example, improving the access for the removal of spare parts.

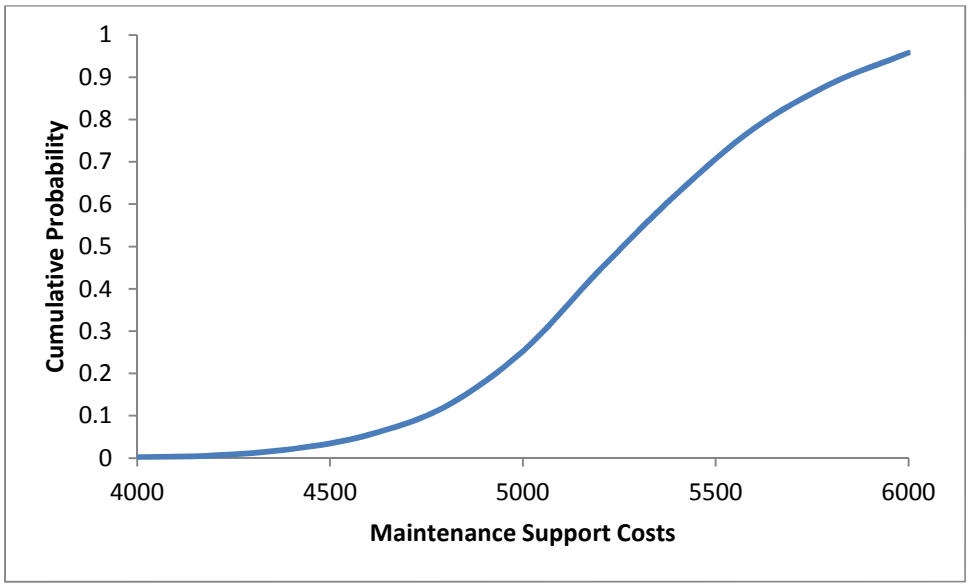


Figure 9 – Plot of the cumulative probability for total maintenance support costs for the example FP.

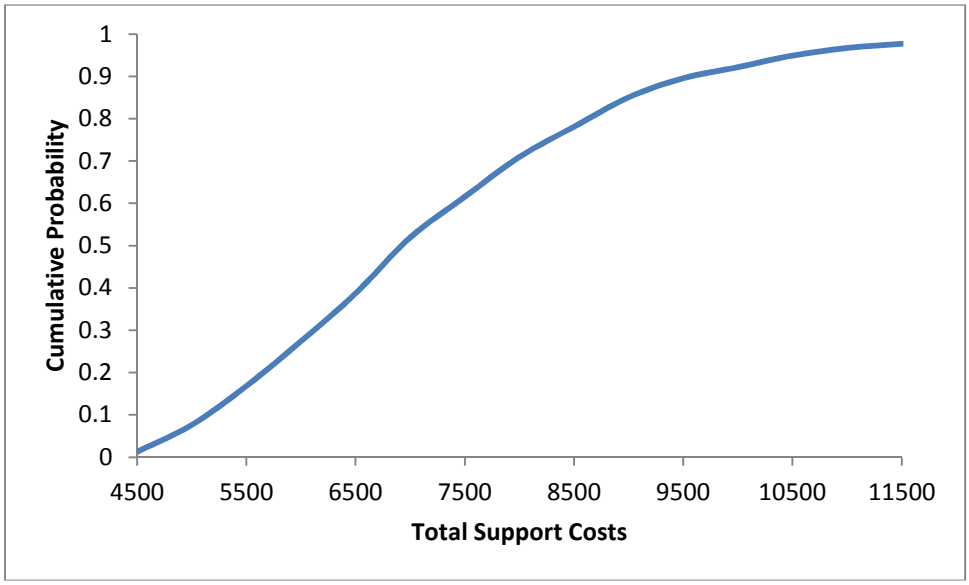


Figure 10 – Plot of the cumulative probability for total support costs for the example FP.

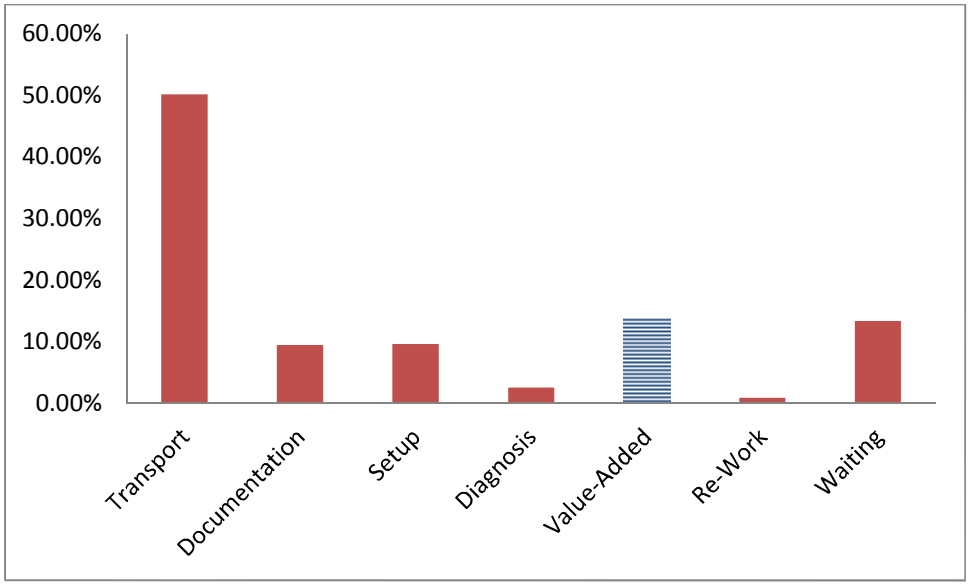


Figure 11 – Estimate of the expected percentage of maintenance engineer hours spent performing different task types.

The efficiency of the support system, calculated through Eqn.

(4), is 0.139 – i.e. 13.9% of the time is spent performing value-added work.

The contributions of the reliability of the components from the example FP to the expected total support costs, calculated through Eqn.

(5), are shown in

Table 4 along with their relative rankings. This shows that failures of the electric motor have the greatest contribution to the expected total support costs, whilst failures of the cooling system contribute the least. However, there is not a substantial difference between their quantitative values.

Component	Hydraulic Pump A	Hydraulic Pump B	Hydraulic Motor	Electric Motor	Cooling System
Contribution (Ranking)	0.36 (=3rd)	0.36 (=3rd)	0.39 (2nd)	0.44 (1st)	0.34 (4th)

Table 4 – Contribution of component failures to expected costs.

The contribution of the completion time for a selection of the maintenance tasks from the example FP (see

Table 2) to the expected total support costs, calculated through Eqn.

(6), are shown in **Error! Reference source not found.** along with their relative rankings. This shows that time spent performing task 1 - travel between the support base and customer site - has the greatest contribution to the total support costs,

therefore basing the engineers on-site rather than off-site would be worthy of consideration to reduce costs.

Task Number	1	5	10	13	15	18	23
Contribution (Ranking)	0.35 (1^{st})	0.01 (8^{th})	0.06 (3^{rd})	0.04 (5^{th})	0.02 (7^{th})	0.07 (2^{nd})	0.03 (6^{th})

Table 5 – Contribution of task completion time to expected costs.

5. Summary and Future Work

The development of FP designs that are optimised for availability and support costs is critical for FP suppliers. In order to do this, they need to be able to predict the performance of FP designs during product development and determine which areas of the design should be prioritised for improvement. A three step methodology has been presented consisting of: representation of the conceptual model, generation of the simulation model and analysis of the data generated from the model. This methodology is not intended to replace human decision making during product development but instead provide decision support information and analysis to the development engineers, helping them to make better design choices. Several metrics that may be useful in identifying elements of an FP design for improvement have been presented. Finally, the methodology has been applied to an example FP to determine its performance and identify some of the main contributors to that performance. This example showed that the methodology could be valuable to those interested in developing FP. There are several areas for improvement to the methodology that are being pursued by the authors. A modelling language is being developed to enable more detailed representation of the conceptual model and the software tool simultaneously developed to generate the simulation model from this representation. This will allow FP performance to be modelled and analysed in much greater detail than is currently possible. The metrics for FP performance analysis introduced in this paper are very basic and the development of further metrics for decision support is another area for future work.

6. Acknowledgements

The research for this paper was supported by the Swedish Foundation for Strategic Research (through the project SSPI) and the Sweden's Innovation Agency VINNOVA Excellence Centre, The Faste Laboratory. John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is

also Director of The Lloyd's Register Educational Trust[1] Centre for Risk and Reliability Engineering at the University of Nottingham.

7. References

[1] M. Goedkoop, C. van Haler, H. te Riele, and P. Rommers, "Product Service-Systems, ecological and economic basics," *Dutch Ministries of Environment (VROM) and Economic Affairs (EZ) Report*, 1999.

[2] S.-H. Kim, M. a. Cohen, and S. Netessine, "Performance Contracting in After-Sales Service Supply Chains," *Management Science*, vol. 53, no. 12, pp. 1843–1858, Dec. 2007.

[3] J. D. Andrews and B. Moss, *Reliability and Risk Assessment*, 2nd ed. Wiley-Blackwell, 2002.

[4] J. Lindström, M. Löfstrand, M. Karlberg, and L. Karlsson, "A development process for Functional Products: hardware, software, service support system and management of operation," *International Journal of Product Development*, vol. 16, no. 3/4, pp. 284–303, 2012.

[5] W. G. Schneeweiss, *Petri Nets for Reliability Modeling*, 1st ed. LiLoLe-Verlag, 1999.

[6] S. Reed, J. Andrews, S. Dunnett, M. Karlberg, L. Karlsson, and L. Magnus, "Modelling Service Support System Reliability," in *IFAC A-MEST*, 2010.

[7] M. Löfstrand, B. Backe, P. Kyösti, J. Lindström, and S. Reed, "A model for predicting and monitoring industrial system availability," *International Journal of Product Development*.

[8] L. Leemis and S. K. Park, *Discrete Event Simulation: A First Course*, 1st ed. Prentice Hall, 2005, p. 624.

[9] T. Ohno, *Toyota Production System: Beyond Large-Scale Production*, 1st ed. Productivity Press, 1988.

[10] J. P. Womack, D. T. Jones, and D. Roos, *The machine that changed the world*, Second. Simon & Schuster Ltd, 2007.

[11] H. J. Harrington, *Business Process Improvement*. McGraw-Hill Professional, 1991.

[1] The Lloyd's Register Educational Trust (The LRET) is an independent charity working to achieve advances in transportation, science, engineering and technology education, training and research worldwide for the benefit of all.

[12] S. Reed, J. Andrews, and S. Dunnett, “Simulation driven design of functional products : a tool for evaluation of hardware reliability and maintenance,” *International Journal of Product Development*, vol. 18, no. 1, pp. 48–67, 2013.

[13] M. van der Borst, “An overview of PSA importance measures,” *Reliability Engineering & System Safety*, vol. 72, no. 3, pp. 241–245, Jun. 2001.

Modelling the Deferred Impact of Failures when Considering the Availability of Production Systems

Jelena Borisevic and Mark Rogers
GL Noble Denton, Holywell Park, Ashby Road, Loughborough,
Leicestershire, LE11 3GR, UK

Abstract

When considering the availability of a production system, it is likely that some equipment failures will not immediately impact production; that is to say, their impact is 'deferred'. However, should this type of failure not be remedied, it will, in time affect production. The impact of a failure could be deferred for a number of reasons, e.g. transient cooling effects, chemical process dynamics, or the use of buffer storage capacity within a system. A system's ability to 'adsorb' critical equipment failures can have a significant effect on production availability. Often, the deferred impact of a failure is modelled as a time based delay, defining a fixed period of time before the failure has an effect on system performance. This approach is commonly used to model the delayed impact of failures of equipment items such as heaters, chemical injection pumps, etc. However, deferred impact can also be related to the volume of buffer storage available within a system – in this case, when considering the benefit of such storage, using a purely time based approach can be an oversimplification. This is particularly true when the flow rate through a system varies significantly over time, or during the repair process. GL Noble Denton has developed a method whereby the deferred impact of failures can be incorporated into Monte Carlo Reliability Availability and Maintainability (RAM) models using a volume-based approach, in addition to the previously used time-based approach. This has provided significant benefits when modelling failures that are not immediately critical due to storage (e.g. ponds for water handling) and typical examples will be given. In addition, GL Noble Denton has also considered how the deferred impact of equipment failures can be restored following a failure and this will also be discussed. The paper also includes a case study to illustrate how the methods have been applied.

1. Introduction

When considering the availability of a production system, the impact of equipment failures, in terms of how they affect overall system performance, needs to be understood. It is likely that some failures will not immediately impact production, that is to say, their impact is 'deferred'. In such a case, if this type of failure is not remedied, it will in time affect production.

The impact of an equipment item failure could be deferred for a number of reasons e.g. transient cooling effects, chemical process dynamics, or the use of buffer storage capacity within a system. A system's ability to 'adsorb' critical equipment failures can have a significant effect on production availability, particularly if the time before the failure becomes critical is of the same order as the repair time. Moreover, a significantly long deferred impact can result in an equipment item having minimal effect on system availability.

In addition to component failures, deferred impact is also used to model the deferred impact of subsystem failures. For example, a loss of production at an offshore platform might not be noticed for a number of hours at an onshore terminal due to drawdown of the linepack in the export pipeline that connects the platform to the terminal. Moreover, if the production outage at the offshore platform is relatively short in duration, it might not even be noticed at the onshore terminal. In addition, there is usually a finite time before the deferred impact is recovered.

There are a number of ways that deferred impact can be modelled. Often, the deferred impact of a failure is modelled as a time based delay, defining a fixed period of time before the failure has an effect on system performance. However, deferred impact can also be related to the volume of buffer storage available within a system – in this case, when considering the benefit of such storage, using a purely time based approach can be an oversimplification. This is particularly true when the flow rate through a system varies significantly over time, or during the repair process.

GL Noble Denton has developed a method whereby the deferred impact of failures can be incorporated into Monte Carlo Reliability Availability and Maintainability (RAM) models using a volume-based approach, in addition to the previously used time-based approach. This has provided significant benefits when modelling failures that are not immediately critical due to the provision of buffer storage.

This paper is divided into three sections:

- Section 2 provides an overview of the different types of deferred impact of a failure on system performance and provides illustrative examples on a single component failure level.
- Section 3 considers the implementation of these techniques in the GL Noble Denton RAM software package OPTAGON [1] and highlights the differences in their application.
- Section 4 discusses the technique and implementation in the case study, with a summary given in the final section.

2. Deferred Impact Types

2.1 Time Based Deferred Impact

The time based deferred impact can be used for the production system components or subsystems whose failure would only affect system performance after a particular period of time had elapsed. An example of this could be an export pipeline corrosion inhibition system on an offshore platform; a failure to inject corrosion inhibitor would not result in an immediate loss of corrosion protection as a film develops on the pipeline surface that has a finite persistency. Other examples might be related to transient cooling effects or chemical process dynamics, which result in equipment failures not immediately impacting production.

Two potential scenarios of such a failure or maintenance impact on the production system can be considered.

Scenario 1: The component or subsystem repair time or maintenance down time is longer than the period of time after which this failure or maintenance would affect the system performance (i.e. deferred impact). This case is illustrated in Figure 1 in a single component model. The grey area in this figure shows the time period when the component is in a fully working condition and, hence, the production is 100% available. The component failure occurs at time T_f, however the production continues due to the deferred impact of the failure for an additional T_{di} period (shown as a dotted area) and stops at the time T_f+T_{di}. The component repair starts immediately after the failure occurrence and takes a time period T_r to restore the component to its working condition. Given the repair time (T_r) is longer than the deferred impact time (T_{di}), the production is unavailable for the time period $T_u = T_r – T_{di}$. Once the component repair is completed the production continues as normal.

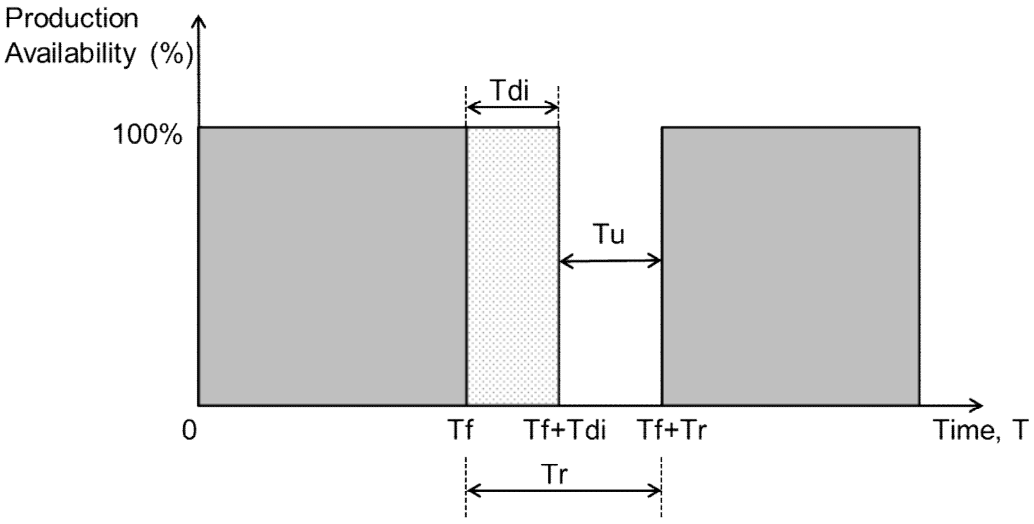


Figure 1. Time Based Deferred Impact Example (Scenario 1)

Scenario 2: The subsequent repair or maintenance of the component or subsystem is completed before the deferred impact time has elapsed. In this case, the component or subsystem operates continues to operate without disruption (although a repair was carried out). This scenario is illustrated in Figure 2 on a single component level. Similar to Figure 1, the grey area on the graph corresponds to the time period when the component is in a fully working condition and the production is 100% available. The component fails at time T_f and is repaired within the time period T_r. The dotted area on the graph indicates that the production doesn't stop during the component repair period (T_r) since the deferred impact period T_{di} is longer than the component repair time. The system is back to the normal working condition at the time T_f+T_r, once the component repair is finished and before the failure becomes critical to system production.

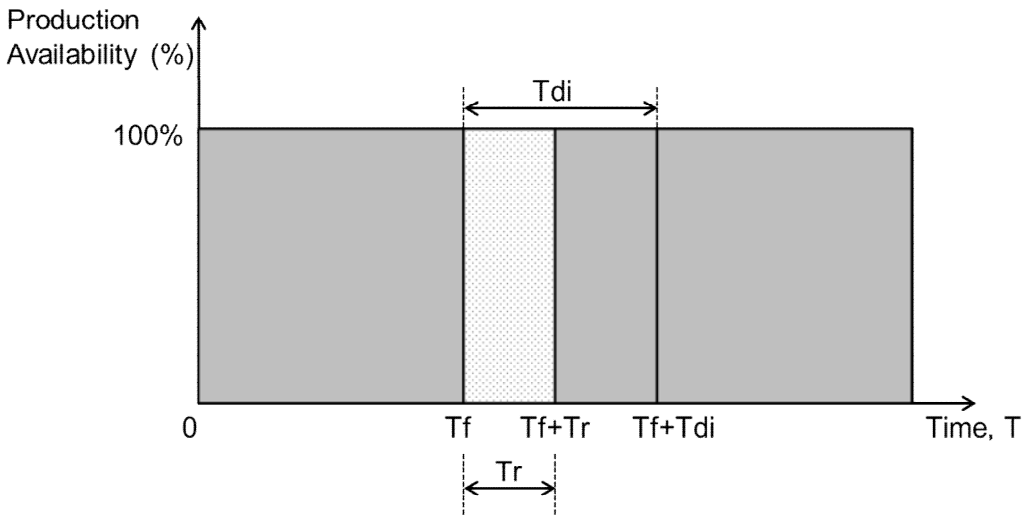


Figure 2. Time Based Deferred Impact Example (Scenario 2)

For example, consider a simple production system that consists of a single component, in which the component fails on average once a year for 12 hours. Assuming exponential distributions for the failure and repair of the component the resulting production unavailability of the system in the absence of any deferred impact is 0.137% (i.e. 12/(8760+12)). If for example, there is a deferred impact of 12 hours before a failure affects system performance, then those failures that are complete in 12 or less hours will not affect the production availability of the system. Assuming an exponential distribution for repairs, this results in approximately 63% of repairs being completed before the failure becomes critical and the production unavailability of the system is reduced to 0.050%. The effect of other deferred impacts for the same example are shown below in Table 1.

Case No.	Deferred Impact (hours)	Production Unavailability (%)
1	0	0.137
2	6	0.083
3	12	0.050
4	24	0.019
5	36	0.007
6	48	0.002
7	72	0.000

Table 1. Example System Production Unavailability for Different Deferred Impact Durations

Table 1 indicates that the deferred impact duration of a component can have a significant effect on the contribution of that component to production unavailability.

2.2 Volume Based Deferred Impact

The time based deferred impact approach described in Section 2.1 can be applied to almost any individual equipment failure or maintenance. However, in some cases, where the deferred impact is provided by a buffer storage volume, the exact duration of deferral may be difficult to define. A typical example would be the line pack of an offshore pipeline. In such a case, a time to empty and a time to re-fill would define the deferred impact on system production. These parameters may vary with time (as production rates change) or in certain operational conditions. Therefore, the application of the time-based deferred impact could be an oversimplification for such failures, particularly for a complex production system.

Consider a simple production system where the liquid product flows into a storage facility (a pond in this case) and is pumped out of the pond either by using a single 100% capacity pump or by using two 50% capacity pumps as shown in Figure 3 and Figure 4 respectively. It is assumed that the pumps serve to maintain the pond below a particular level (e.g. 50% full) and that normally the inflow equals the outflow (i.e. at a 100% rate).

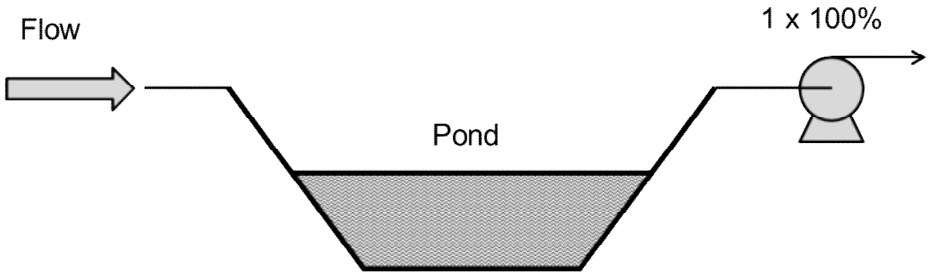


Figure 3. Example Production System (Single Pump)

If the downstream pump failure occurs in the 1x100% case (Figure 3), the pond will start to fill and the deferred impact on the system production will be defined by the upstream flow and the available pond volume.

Alternatively, if one of the 2 x 50% downstream pumps fails (Figure 4), the deferred impact duration will be twice as long as in the 1x100% case. Once the pond is full, system production can continue but only at 50% production level.

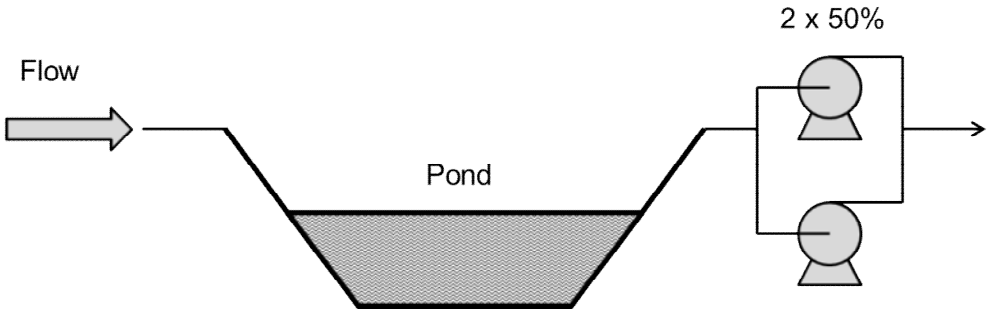


Figure 4. Example Production System (Two Pumps)

This example shows that the volume based deferred impact has a longer effect if some outflow is maintained. It is also found that a volume based deferred impact is particularly important when the flow varies significantly with time or even during the repair. Therefore, the volume based deferred impact modelling approach should be used in such cases.

3. Deferred Impact Modelling with OPTAGON

3.1 OPTAGON Overview

GL Noble Denton has developed a Monte Carlo simulation tool OPTAGON [1] for predicting and optimising the availability of complex oil, gas and water production facilities [2]-[6].

When using OPTAGON, the system to be investigated is described by a reliability block diagram that shows how the operation of the system as a whole depends on its critical components. These components are arranged in series and parallel groups to form an overall representation of the operation of the system. Components are arranged in series groups if they are all required to work for that group to be operational and in parallel groups if there is some redundancy within the group of components. Series and parallel groups can be nested to describe the operation of the system. A simple OPTAGON model for the example production system from Figure 4 is shown in Figure 5 below.

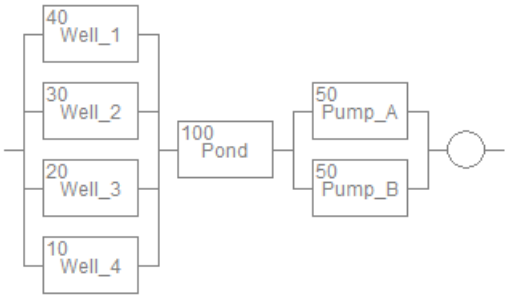


Figure 5: Example OPTAGON Model

This concept is extended in OPTAGON by associating a capacity, or flow, with each component. The capacity of a series group of components is the smallest capacity of any of its components, and the capacity of each parallel group is the sum of the capacities of its components. The complete model has a system capacity depending on the configuration that can vary with time if component capacities (such as well components) vary over time. By using associating capacities with components, the reliability block diagram can represent partial operation of a system.

A demand production rate is placed on the model. This demand can be equal to the system capacity or, as in often the case in gas production systems, can

be a separate demand profile that varies with time and is slightly below the maximum system capacity.

Each component in the reliability block representation can have a wide variety of attributes including failures and repair characteristics, varying process throughput, preventative maintenance schedules, spare requirements for repairs, type of redundancy and logistic delays.

3.2 Monte Carlo Simulation Technique Implementation

Results in OPTAGON are generated by using the Monte Carlo simulation technique [7], a stochastic method for modelling complex systems, where the behaviour is too subtle or sophisticated to be modelled accurately (or at all) by analytical methods. The technique consists of explicitly modelling the production system being studied, subjecting it to a typical set of events over its lifetime, and empirically observing how well it performs. The typical events which are directed at the model are generated stochastically, through the use of effectively random numbers. This means that any individual simulation of the system's lifetime cannot be taken as a guide to its average performance since the model may have been subject to events that were more or less favourable than the average. Instead, it is necessary to carry out a large number of individual simulations. The performance of the model over many simulations gives an indication of how it is likely to perform on average, and how widely the range of possible performance is spread.

The stochastic nature of the simulation arises from the way in which the times for the components to fail and to be repaired are specified and generated. The explicit nature of the model allows including a wide variety of complex component and system behaviours, such as the use of the deferred impact, without having to reduce them to analytical approximations.

It is in the nature of event driven simulation that each simulated lifetime of the model is different, and the stream of random numbers used to generate the events will have different meanings in each case. The differing values of random numbers in each run mean that events happen in different sequences, and a random number which is used in one run to generate a failure time may be used in the next to generate a repair time.

3.3 Deferred Impact Modelling

OPTAGON allows modelling both the time based and the volume based deferred impacts on the production system:

Time Based Deferred Impact Modelling: when modelling the time based deferred impact in OPTAGON, the time to become critical and the time, following repair, for this non-criticality to be regained are defined (at either component level or group level). A pro rata recovery time option is also available. In pro rata case, the recovery time is proportional to the used deferred impact. For example, if the maximum recovery time is 24 hours and

6 hours of 9 hours of deferred impact have been used, the recovery time is 16 hours. Conversely, if all the deferred impact has been used, the recovery time would be 24 hours.

Volume Based Deferred Impact Modelling: Alternatively to the time based approach, a buffer volume can be specified which, together with the flow through the component at the time of the failure, determines the time to become critical. In this case, a lower level of the flow through the component when the failure occurs, results in the deferred impact lasting longer. If the flow through the component changes when the deferred impact is being used (either due to capacity changes or other failures/repairs), the deferred impact changes accordingly.

3.4 Example

In order to illustrate why it is important to be have the flexibility to model deferred impact using a volume based method, in addition to a time-based method, a relatively simple system can be considered. The system can be represented by a single component which fails every 6 months (i.e. exponential distribution with a Mean Time Between Failures (MTBF) is equal to 4380 hours) for half a day (i.e. exponential distribution with a Mean Down Time (MDT) is equal to 12 hours). The system can provide a 1000Mscf buffer volume when such a failure occurs. The gas flow is changing during the 5 years production period as shown in Table 2 below. Table 2 also summarises the calculated volume based and the time based deferred impact for each production year. It can be noticed that the volume based deferred impact on the production system changes annually as it is calculated from the actual gas flow relative to the available buffer volume. The time-based deferred impact is based on the average gas flow of 44Mscf/hour and, therefore, is constant throughout the 5 years production period.

Production Year	Gas Flow Rate (Mscf/hour)	MTBF (hours)	MDT (hours)	Buffer Volume (Mscf)	Volume Based Deferred Impact (hours)	Time Based Deferred Impact (hours)
1	100	4380	12	1000	10	23
2	50	4380	12	1000	20	23
3	25	4380	12	1000	40	23
4	25	4380	12	1000	40	23
5	20	4380	12	1000	50	23

Table 2. Deferred Impact Comparison

Figure 6 compares the buffer storage available for the production system following its component or subsystem failure by using the time based (shown

in grey solid line) and the volume based (represented by black solid line) deferred impact modelling approaches.

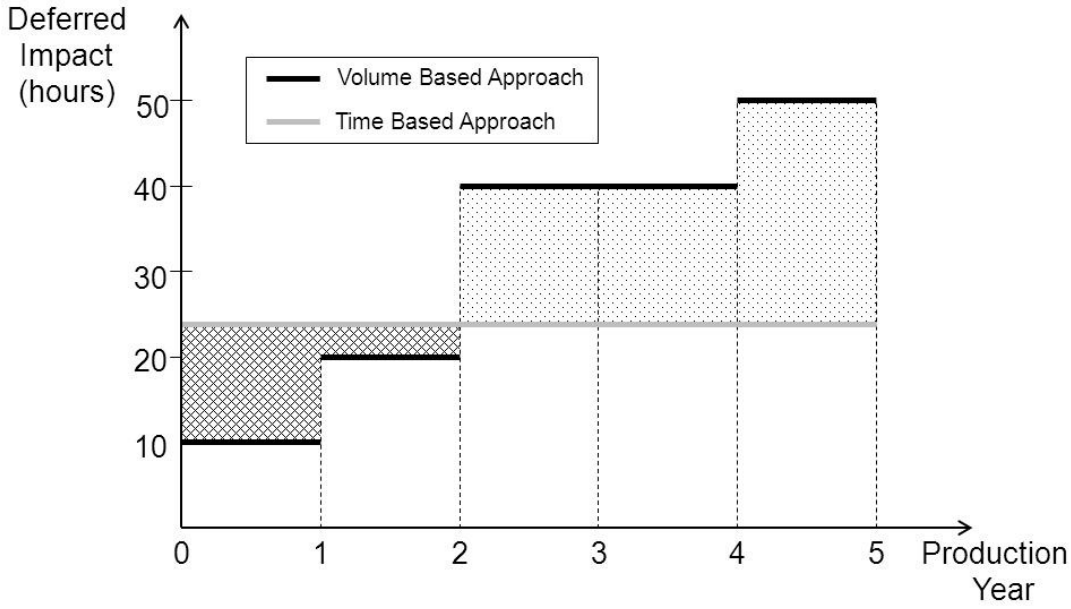


Figure 6: Example Production System Annual Buffer Storage Comparison

Figure 6 shows that during the first two years of gas production the system deferred impact prediction using the time based approach (represented by outlined Diamond area on the graph) is optimistic in comparison to the deferred impact modelled explicitly using the volume based approach. Conversely, during the last three years of production the time based approach underestimates the available deferred impact (shown as a dotted area on the graph).

Figure 7 compares the resulting annual system production availability calculated by using the volume based deferred impact approach (highlighted by a black solid line) to the system production availability obtained by using the time based approach (represented by grey solid line).

Considering the 5-year period in total, the volume-based method indicated an overall average production availability of 99.93% compared to the optimistic 99.96% using the time-based method.

It can be seen in Figure 7 that the time based deferred impact modelling approach provides only an average, and in this case optimistic, estimate of the system production availability. Conversely, the volume based approach provides a detailed and more realistic estimate of the system performance, which is very important when modelling a complex production system.

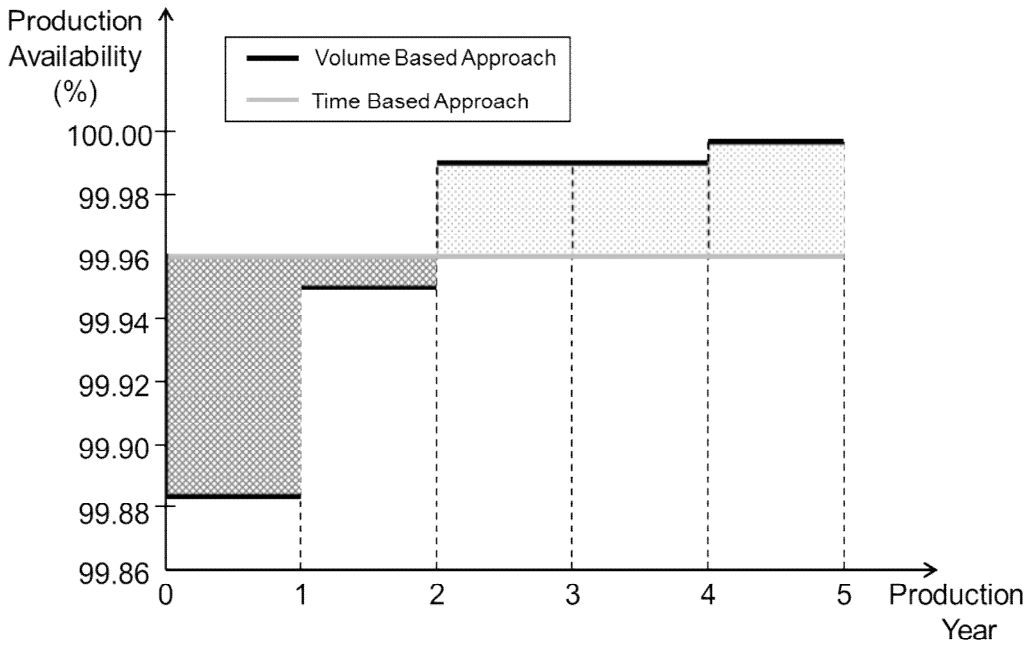


Figure 7. Example Production System Annual Production Availability Comparison

4. Case Study Implementation

The concept of the volume based deferred impact modelling discussed in previous sections was recently successfully applied by GL Noble Denton in a complex RAM study for one of the major coal seam gas and to LNG development projects in Australia. In order to assess gas deliverability, GL Noble Denton was asked to carry out a RAM assessment by developing an integrated model for the upstream facilities, comprising

- Gas gathering, processing and transportation facilities
- Water gathering, storage and treatment facilities
- Power and communication systems
- The impact of LNG plant availability on the upstream facilities.

Figure 8 shows a simplified schematic of the water gathering system considered in the RAM study. The water management facilities included water gathering and regional storage ponds, together with pumps that transport water via water trunklines to raw water storage ponds upstream of the Water Treatment Plants. A Water Treatment Plant could receive produced water from a number of regional storage ponds. The operating philosophy was to operate regional ponds at a minimum level to cater for downstream outages, while the raw water ponds would fluctuate depending on operational variability. The working pond volume was based on a percentage of the total pond volume estimated using the operational trigger levels for the pond pumps coming on and switching off.

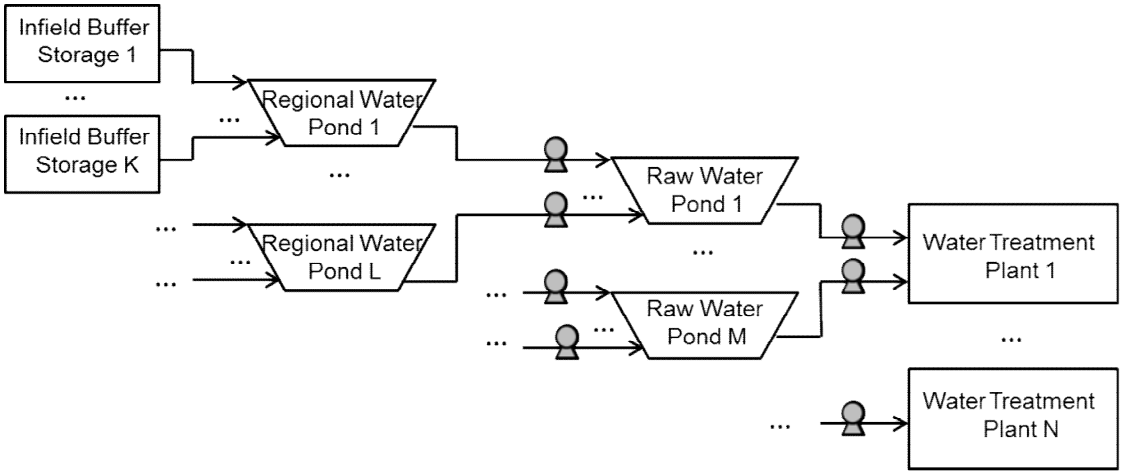


Figure 8: Water Gathering System Schematic

A number of different pond pumping arrangements were considered, including the following:

- *Pond served by a single pond pump (i.e. 1 x 100% arrangement)*: if the pump (or associated equipment) failed, then the spare capacity within the pond was filled at the current water flow rate for that pond. If the failure was not repaired before the pond became full (i.e. buffer volume completely used up), then the gas flow from the facility served by that pond would have to be switched off. Once, the repair was complete the gas flow could be turned back on immediately and it was assumed that it would take two weeks to restore the normal spare capacity in the regional pond (this time was considered conservative and was likely to be shorter). If the failure was repaired before the pond was full, the time to restore the normal spare capacity was pro-rated against the amount of capacity filled. For example, if the repair was completed when half the spare capacity had been used, it would take 1 week (i.e. 50% of 2 weeks) to restore the spare capacity.

- *Pond served by two pumps (i.e. 2 x 50% arrangement)*: Given such an arrangement, if one pump failed then the spare capacity within the pond would take twice as long to fill compared to a 1 x 100% arrangement. Moreover, if the repair was not undertaken before the pond was full, it was assumed that the gas flow rate would fall to 50% when the pond was full (as one pump is still operational). If a failure occurred such that both pumps could not operate (e.g. complete power failure at location), then the filling of the pond would be the same as in the 1 x 100% arrangement described above.

The filling rates assumed for the various ponds in the water gathering system varied significantly over time; therefore it was essential that the ponds were modelled with a volume-based deferred effect. It was found there was only a very small likelihood of ponds becoming full and causing gas production to be shutdown from wells. This was found for both regional ponds and raw water

ponds. This was a significant result given initial concerns regarding insufficient upstream buffer storage.

Although the deferred impact of the downstream failures was modelled using the volume-based deferred effect to represent the filling of ponds, the subsequent emptying of ponds back to normal levels following repair completion was modelled using a time-based approach. A sensitivity analysis on the time to empty the pond was undertaken and it was found that the results were insensitive to a wide range of emptying times (larger and smaller than the two week period used). This result was directly related to the low probability of a significant downstream failure (causing a pond to start filling above its normal level and becoming full) happening very soon after a previous significant downstream failure.

5. Summary

This paper has described the need to include deferred impact of failures when considering the availability of a production system. The limitations of modelling the delayed of some failures using a 'time-based' method have been outlined and another technique, based on a 'volume-based method' has been described. This 'volume-based' method has been developed by GL Noble Denton for inclusion in its OPTAGON RAM software package and provides significant benefits when modelling failures that are not immediately critical due to storage, particularly when production rates vary significantly with time over the period being considered. A case study has provided an example of this volume based approach.

References

1. www.OPTAGON.info
2. Rogers M. and Johnson M., *Developing the OPTAGON Package to Enable Availability Modelling of Challenging Gas Production Scenarios*, Advantica Technology, 15th AR2TS (2003).
3. Trim S., Coote M. and Rogers M., *Effective Assessment of Potential System Changes to Optimise the Value of an Asset*, ERTC (2004).
4. Wragg N., *optimising Design and Maximising Performance of complex Integrated Asset Systems Using OPTAGON*, Advantica Ltd, ERTC (2008).
5. Arizmendi-Sanchez J. and Wragg N., *Exploiting Benefits from Expanded Applications of reliability Modelling with OPTAGON Software*, GL Industrial Services, ERTC (2009).
6. Minnit J., *OPTAGON: Pushing the Boundaries – Cann RAM Modelling Technique be Applied to more Complex Operations?*, GL Noble Denton, ERTC (2010).
7. Tanner M. A., *Tools for Statistical Inference: Methods for the Exploration of Posterior Distribution and Likelihood Functions*, Springer Series in Statistics, 3rd Edition, Springer Verlag (1996).

A Markov Modelling Approach to Railway Bridge Asset Management

Bryant Le and John Andrews
Nottingham Transportation Engineering Centre, University of Nottingham

Abstract

The UK has a long history in the railway industry with a large number of railway assets. Railway bridges form one of the major asset groups with more than 35,000 bridges. The majority of the bridge population are old being constructed over 100 years ago. Each bridge is a unique construction featuring different structure arrangements and component materials. The decision of when and what type of maintenance actions should be performed is a difficult task. Moreover, the problem becomes more complex as other factors such as the route criticality and possession time affects the maintenance planning. Models can be formulated to predict the future condition of assets along with the effect that interventions such as servicing, repair and replacement will produce, these can be used to support this decision making process.

This study demonstrates a Markov modelling approach to predict the condition of individual bridge elements. For each bridge element the deterioration process is determined by examining historical maintenance work and analysing the times that each element takes to deteriorate to the point where maintenance of a certain severity classification is required. A complete bridge model is also constructed based on these elemental models. The bridge is considered in terms of individual elements such as: decks, girders, abutments and bearings and the focus of the study is the asset group of metal underbridges.

1 Introduction

Railway bridges in the UK are categorised into underbridges and overbridges. Overbridges carry another service over the railway, while underbridges carry rail traffic over obstacles. A significant proportion of the underbridge population are metal bridges and the majority of these were constructed before 1912. The load carrying by underbridges are often much higher than the load carried by overbridges because the load imposed by rail traffic is much more than road vehicles. Due to the fact that the current network demand is increasing, the traffic loads and intensities on the structures are also increasing making underbridges one of the most critical part in the bridge asset group. Hence the focus of this paper is studying the metal railway

underbridges. A bridge is a complex structure with components of different materials and many bridge models in the literature consider the bridge as a whole, this reduces the accuracy of the model. A bridge in this paper is considered in term of individual principle elements such as: decks, girders, abutments and bearings. These elements are also categorised according to their material: metal, concrete, timber or masonry. In order to make sound decisions regarding when and how to maintain the structure, the deterioration process of the bridge component over time must be well understood. Maintenance costs can then be incorporated into the model allowing the projected maintenance costs over the life of the asset to be investigated. The whole life cycle costs (WLCC) of different maintenance strategies can also be studied and compared with each other.

There are a few approaches to model the degradation process of bridges. The Markov modelling approach has been widely used for this purpose. Several models have been developed to predict the future condition of an asset for use to support maintenance decisions. These models can be classified as Markov, semi-Markov and probabilistic models.

Jiang et al [1] and Robelin & Madanat [2] explained the use of Markov models in predicting the deterioration rate of bridges. The deterioration rate of the bridge is reflected in the Markov state transition probability matrix where each element of the matrix represents the probability of the bridge moving to a certain condition. Based on the condition score data of bridges at different ages, the state transition probability matrix was estimated by minimising the absolute difference between the expected value of the condition rating from the Markov chain and the actual average condition rating from the database. Cesare et al [3], Ortiz-García et al [4] and Chase & Gaspa [5] presented real applications of Markov models to the evaluation of bridge deterioration. The studies were carried out based on the data of bridges in different states in the US. The data contains bridge element condition ratings on a scale from 1 to 7, with 7 being ‘as new’ and 1 being the worst condition. The Markov model was then applied to predict the evolution of the average condition rating of a set of bridges and the expected value of the condition rating for a single bridge. Morcous [6] investigated the effect of a non-constant inspection period on the transition matrix and proposed that the transition probabilities should be updated using Bayes’ rules for more accurate modelling.

While the Markov model is based on the assumption of an exponential distribution for duration (sojourn) times in specific bridge conditions, semi-Markov models use different distributions (often the Weibull distribution) to model these duration times.

Ng and Moses [7] discussed the use of semi-Markov processes in modeling bridge deterioration. Each state’s sojourn time distribution parameters were

estimated using the difference between the two age distributions for the respective states. The study was based on real bridge condition data, however, the condition data is for the whole bridge, not bridge elements. In Kleiner [8], the author discusses modelling the waiting time of the process in any state as a random variable with a two-parameter Weibull probability distribution. The application of the model is then demonstrated based on hypothetical data, which was obtained from expert opinion and perception. Having the Weibull distribution parameters defined by the experts, the transition matrix was then obtained and the future condition of the assets was predicted. Empirical models which used real condition data can be found in Sobanjo et al [9], Mishlani and Madanat [10] and Yang et al [11]. Weibull distributions were fitted to the times of a bridge component remaining in a particular condition rating. The transition probability between states can be calculated at any point in time employing a semi-Markov model to account for the non-constant deterioration rate of the bridge element.

Probabilistic models were developed by Agrawal et al. [12], where the author studied 17,000 highway bridges in New York State with historical data available from 1981 to 2008. Again the bridge component condition ratings were on a scale from 1 to 7. The approach fitted a Weibull distribution to the durations that a bridge element stays in a particular condition then calculates the mean time of staying in that particular state. The mean duration for each different condition rating is calculated by accumulating the mean durations of the previous states. These means are then plotted on a graph of condition ratings against age and a third degree polynomial fitted to show the deterioration rate. Frangopol et al [13] took a different approach and developed a reliability index that measures the bridge safety instead of condition and the deterioration rate of a bridge is the rate of deterioration of the reliability index.

Markov, Semi-Markov and probabilistic approaches have previously been applied to bridge assessments. There are however limitations in these models such as: their basis on condition rating data which is not an ideal for the determination of degradation processes or maintenance models, the estimation of the transition probabilities is significantly affected by prior maintenance actions (i.e. a rise in condition score) (Agrawal et al [12]); the effects of maintenance on components are not captured.

The bridge model in this paper addresses these deficiencies by the use of historical maintenance data instead of condition data. This gives the time to an event when an intervention was carried out. The model is a Markov model that represents the life of bridge components taking account of their current condition, material, structural type and environment. The model is also capable of accounting for the maintenance strategy, inspection interval, servicing interval and the repair delay time.

2 Bridge element conditions and intervention types

Over time, the bridge element condition deteriorates and structural defects appear which trigger different types of intervention. Different components of the bridge are constructed from different materials and would experience different types of degradation. Thus, the maintenance actions required by each component of the bridge would be different. When the component reaches a certain level of defect, a certain type of maintenance is triggered. There are four intervention categories considered which are given in Table 1. Servicing is the only type of maintenance which does not change the state of the component, servicing will slow down the degradation rate. Minor repair, major repair and replacement are assumed to restore the component condition to as good as new. These three interventions can be carried out when the component reaches the good, poor or very poor state from the 'as new' condition.

Maintenance type	**Definition**			
Servicing	Activities that protect the structure from the source that drives the degradation process (eg. Waterproofing).			
Minor repair	Minor repair implies the restoration of the structure element from the good condition to the as new condition. Components in the good condition can experience the following defects			
	Metal	*Concrete*	*Timber*	*Masonry*
	Minor corrosion, tear	Spalling, small cracks, exposed of secondary reinforcement	Surface softening, splits	Spalling, pointing degradation water ingress
Major repair	Major repair implies the restoration of the structure element from the poor condition to the as new condition. Components in the poor condition can experience the following defects			
	Metal	*Concrete*	*Timber*	*Masonry*
	Major corrosion, loss of section, fracture, crack welds	Exposed of primary reinforcement	Surface and internal softening, crushing, loss of timber section	Spalling, hollowness, drumming
Replacement	Complete replacement of a component or the whole bridge. Components in the very poor condition can experience the following defects			
	Metal	*Concrete*	*Timber*	*Masonry*
	Major loss of section, buckling, permanent distortion	Permanent structural damage	Permanent structural damage	Missing masonry, permanent distortion

Table 1: Intervention categories based on the level of defects for different bridge element materials

3 Analysis of bridge element deterioration

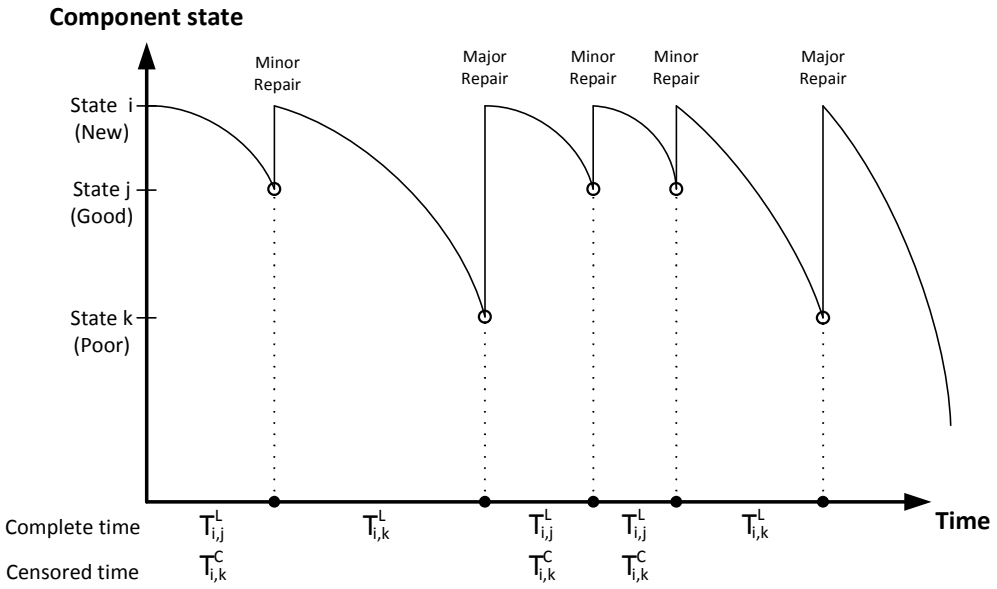


Figure 1: History of work on a component

From the historical maintenance records, for each bridge component, a history of work done on the component can be constructed as illustrated in Figure 1. The degradation process is determined by analysing the times that each element takes to deteriorate to the point where a certain type of maintenance is required. The time to reach state j (good) and k (poor) from new (state i) are given as $T_{i,j}^{L}$ and $T_{i,k}^{L}$. These times are often called the time to failure and the term 'failure' used here does not mean the physical failure of a component but indicates the time to the point when a certain type of repair is necessary. It is important when analysing the lifetime data of a component to account for both complete data, $T_{i,k}^{L}$ and censored data, $T_{i,k}^{C}$. Complete data indicates the time of reaching state k from the new condition. Censored data is incomplete data where it has not been possible to measure the full lifetime. This may be because the component was repaired or replaced, for some reason, prior to reaching the condition *k* and so the full life has not been observed. The components life is however known to be at least $T_{i,k}^{C}$. Figure 1 shows how the complete and censored times have been analysed. The time between any repair and the next minor repair is a complete time indicating the full life time of the component reaching the state where minor repair is required from new state. This time is also the censored time recording for the time to major repair needed. The component's condition was restored to new condition before it reaches the state where major repair is necessary hence the full time to a major repair cannot be measured. Having obtained these data for a single component, the analysis can be performed on similar components of the same type and material. The transition rates, λ_j between the new state to state *j* can be estimated using Equation 1.

$$\lambda_j = \frac{\sum_1^n N_i}{\sum_1^n \left[T_{i,j}^L + T_{i,j}^C\right]} \quad \textbf{(1)}$$

where
Ni is the number of repairs on a single component
n is the number of component of type *i* studied

Assuming that the degradation rate is constant, the mean time to an intervention of type *j* (MTTI) can be calculated as the inverse of the degradation rate. The deterioration rates for the four different main bridge components of different materials are presented in Table 2. Considering the bridge deck, it shows that, the rate of timber deckings reaching a point where they would require a minor repair is considerably higher than the corresponding rates of metal and concrete decking. This indicates that a timber deck would need more maintenance than the metal and concrete deckings. While the average replacement rate of timber decking is after 45 years, metal decking life is about three times longer (136 years). For concrete decking, it is only needed to be replaced after 187 years, it has the longest working life, 4 times longer than timber deckings and 1.5 times more than that of metal. The study shows that the bearing deteriorates faster once it reaches the good condition. Note that there were not enough data to allow the replacement rate of the abutment to be estimated, this also suggests that bridge abutments are rarely replaced.

Bridge Component	Material	Condition	Intervention	λ (year^{-1})	MTTI (year)
GIRDER	Metal	Good	Minor Repair	0.03666	27.28
		Poor	Major Repair	0.01845	54.21
		Very Poor	Replacement	0.00461	216.84
DECK	Metal	Good	Minor Repair	0.01591	62.85
		Poor	Major Repair	0.01136	88.00
		Very Poor	Replacement	0.00734	136.32
	Concrete	Good	Minor Repair	0.01327	75.34
		Poor	Major Repair	0.00733	136.43
		Very Poor	Replacement	0.00533	187.51
	Timber	Good	Minor Repair	0.10885	9.19
		Poor	Major Repair	0.03173	31.52
		Very Poor	Replacement	0.02212	45.20
BEARING	Metal	Good	Minor Repair	0.02284	43.78
		Poor	Major Repair	0.01845	52.19
		Very Poor	Replacement	0.00461	67.18
ABUTMENT	Masonry	Good	Minor Repair	0.01925	51.94
		Poor	Major Repair	0.01845	100.87
		Very Poor	Replacement	-	-

Table 2: Deterioration rates and mean times to an intervention (MTTI) of bridge elements

4 Bridge model

This section contains a description of the construction of a continuous-time Markov model that models the degradation, inspection, servicing and maintenance of the all bridge components. The models for the bridge elements as well as the whole bridge are presented. Analysis can then be carried out on the complete bridge model to investigate the effects of different maintenance strategies.

4.1 Degradation, inspection and maintenance strategy

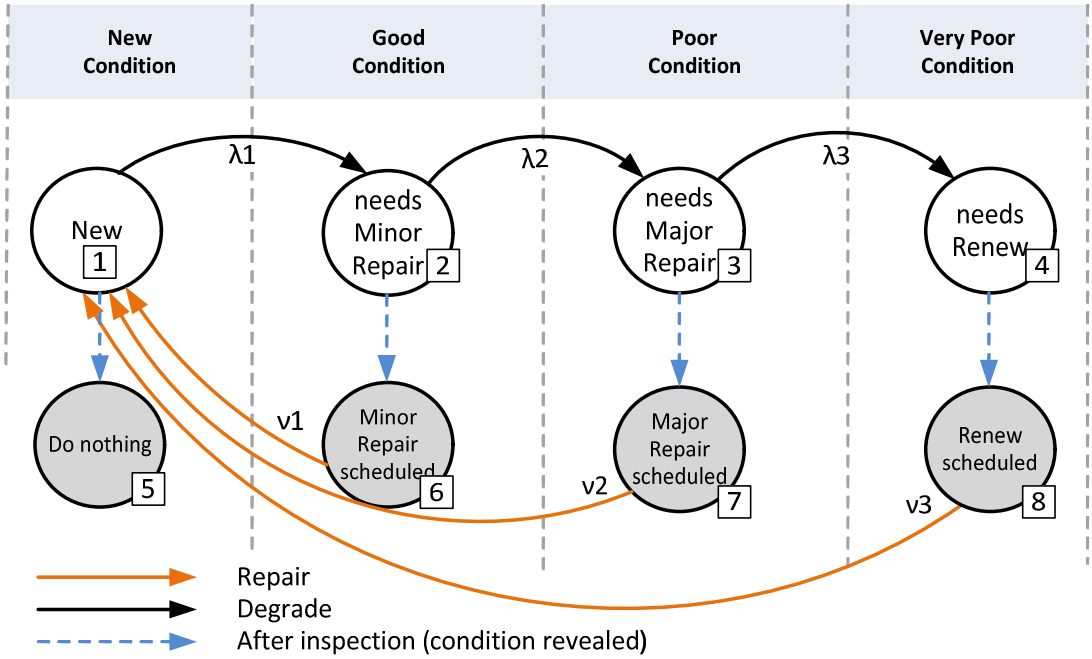


Figure 2: Markov states for a single bridge element

It is assumed that a bridge component can be in one of four conditions as discussed in the previous section. The component starts in the new condition and degrades to the good condition where the component requires minor repair. Further deterioration leads to the poor and very poor conditions where the component requires major repair and complete renewal respectively. All bridges and their components are inspected after a certain period of time. At the point of inspection, the current state of the bridge component is revealed and the decision can be made to repair or let it deteriorate to a poorer state. In the period between two inspections, the true state of the component is unknown, i.e. the actual condition might be worse than the last known condition. Figure 2 shows the Markov state diagram that was developed to model the degradation and repair processes of an element. The component starts in the new condition (State 1) and deteriorates to State 2 where a minor repair can be performed. Following an inspection, if it is revealed that the component is in State 2, the element can either be scheduled for repair (State 6) or left to deteriorate to poorer state. The option to carry out repair is achieved by enabling the repair process represented by the arrow connecting State 6 and State 1. In contrast, if the fore-mentioned arrow is removed, the repair process is disabled, it means that even if the component is discovered

to be in the state where minor repair is possible, no action is being performed and the component continues to deteriorate to further states. A similar process applies when the component deteriorates to a state where a major repair is necessary to return it to the as new condition (State 7), the option for repair or no repair is again set by the repair process represented by an arrow connecting States 7 and 1. Note that State 8 is when the component is revealed to be in a very poor condition and cannot deteriorate any further, since the component should be repaired as soon as it reaches this level of condition the repair process between State 8 and State 1 should always be enabled. Effectively, the model models two phases in the component's life: the first phase is the continuous phase, modelling the degradation and repair processes, between any two inspections and the second phase is at the point of inspection where the condition of a bridge element is revealed and the decision whether to repair or not is made. There are four maintenance strategies possible in this model and are described in Table 3.

Strategy	Action
Strategy 1	Repair as soon as the component is identified to be in a state where repair is necessary, then it is carried out.
Strategy 2	Repair when the component is identified in the state where a major repair is required i.e. repair when the component reaches poor condition.
Strategy 3	Repair when the component is identified as being in the state where renewal is needed i.e. repair when the component reaches very poor condition.
Strategy 4	No repair, component is allowed to deteriorate without any interventions

Table 3: Maintenance strategies possible in the bridge model

4.2 Single bridge model

A bridge contains many components thus extra states must be added to the model to uniquely represent the states of each individual bridge element. It is worth noting that the number of states in a Markov model increases exponentially as the number of components in the model increases. It is not possible to illustrate the complete bridge model graphically, Figure 3 illustrates the states of the Markov model, considering a system of only two components - two main girders. Each bridge main girder can be in four possible conditions, thus there are 4^2=16 possible Markov states. At the point of inspection, the conditions of the components are revealed therefore, an extra 16 states are added to the model representing the states where the component conditions are actually known. In Figure 3, degradation transitions between the states are represented by darker shade solid arrows, repair

transitions are presented by lighter shade solid arrows and inspection transitions are denoted by dashed arrows. The shaded states are the states in which the bridge element conditions are revealed. For example, in state 24, it is revealed that after the inspection, the main girder 1 (G1) is in a good (G) condition while the main girder 2 (G2) is in a very poor (VP) condition. If the maintenance strategy is to repair the component as soon as they reach the states where repair is necessary then the repair process will restore the girders' conditions to as good as new, this is represented by the repair process from state 24 to state 1.

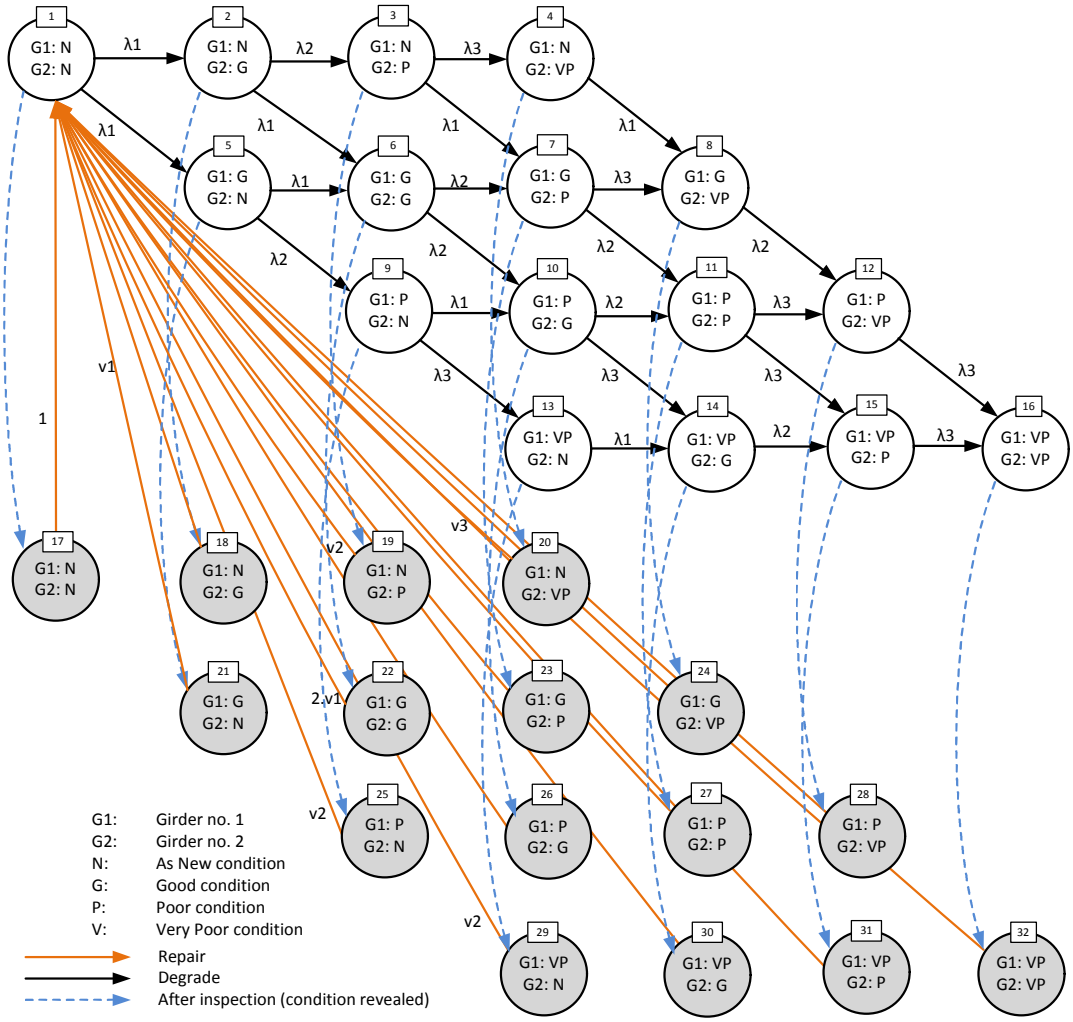


Figure 3: Markov bridge model for two main girders assuming strategy 1 (repair as soon as the repairs are necessary)

The Markov bridge model states are the same for different maintenance strategies however the repair processes are different. Figure 4 illustrates the model for the same two girders system that is managed under maintenance strategy 3. Note that the degrade transitions are the same as illustrated in Figure 3 therefore they are not shown in this figure for the sake of clarity. Computer software was written in Matlab to aid the process of generating the larger and more detailed Markov bridge model to include more components.

The software first generates all of the possible model states and then generates all the transitions possible governed by a specified maintenance strategy.

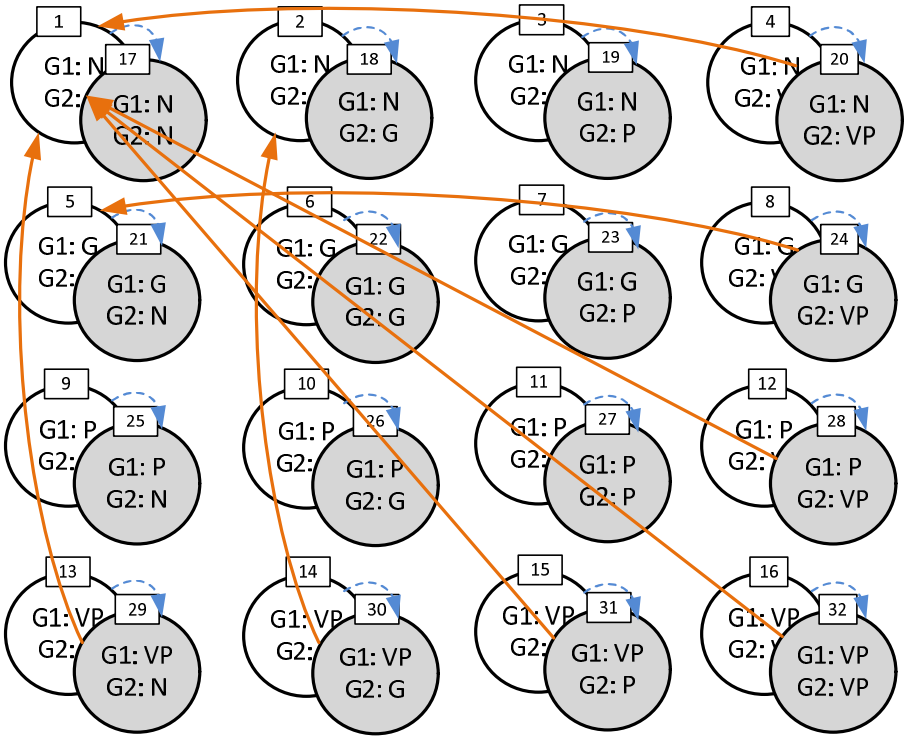


Figure 4: Markov bridge model for two main girders assuming strategy 3 (repair when the component reaches the condition where renewal is needed)

4.3 Opportunistic maintenance

Since the bridge model consists of many different elements, the conditions and the deterioration rates of these elements are different hence the times when interventions are required are also different. If a component is being repaired, opportunistic maintenance considers carrying out repair on the same adjacent component that does not necessarily need the same type repair as the first component but is in the state where repair is possible. Figure 5 shows the model for two main girders with maintenance strategy 2 and identifies the states where opportunistic repair is possible. In particular, state 23 in the model represents the case that after an inspection the main girder 1 is discovered to be in a good condition and the main girder 2 is in a poor condition. Under maintenance strategy 2, only the main girder 2 is being repaired, the repair process will bring the system to state 5 where the main girder 2 is now in the as new condition whilst the main girder 1 remains in the same condition. It is possible in this case to carry out opportunistic repair on the main girder 1, this will bring the system back to state 1 where both component conditions are restored to as good as new. The repair process

represented by an arrow connecting State 23 and State 5 will be replaced by an arrow from State 23 to State 5. Again the process of generating the repair process for opportunistic maintenance is done automatically using the written software.

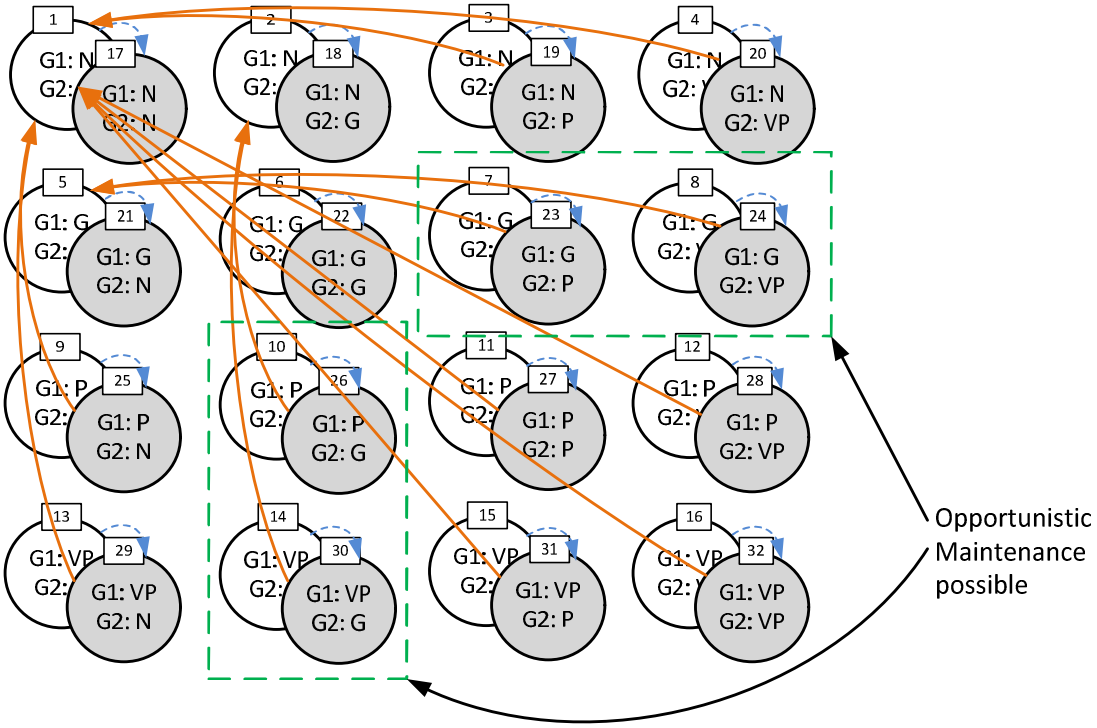


Figure 5: States where opportunistic maintenance are possible in a system consisting of two main girders operating under maintenance strategy 2

4.4 Transition rates

The deterioration rates established above govern the process from the 'as new' state. The Markov model needs the transition rates between two adjacent states (good to poor, poor to very poor). The rate from state *i* to state *j* can be estimated as the inverse of the mean time reaching state *j* from state *i*, $MTTF_{i,j}$:

$$MTTF_{i,j} = MTTF_{1,j} - MTTF_{1,i} \quad \textbf{(2)}$$

Giving:

$$\lambda_{ij} = \frac{\lambda_i \lambda_j}{\lambda_i - \lambda_j} \quad \textbf{(3)}$$

The repair rates, the inverse of the mean time to repair (MTTR), ν_1, ν_2, ν_3 are also included in the model. The time to repair consists of two main components:

- the time to schedule the repair
- the time of the actual repair work being carried out (possession hours)

The schedule time for the work is defined as the duration between when the work was raised and when the work actually starts. The time of repair is calculated as the duration of the work. Thus the repair rate can be calculated as:

$$\nu = \frac{1}{MTTR} \quad \textbf{(4)}$$

As mentioned earlier, there are two phases for the model calculations:

The first phase is the continuous phase between any two inspections, the system equation is governed by equation (5) where *Q* is the matrix representing the probabilities of being in each state; *A* is the transition rate matrix based on the deterioration rates and repair rates; and $\dot{Q}$ is the rate of change of probabilities at each state in the model. This system of differential equations was solved by the 4th order Runge-Kutta method with variable time step to speed up the process. The average step size is 0.03 year.

$$\dot{Q} = Q.A \quad \textbf{(5)}$$

The second phase, corresponding to the point of inspection, is a discrete phase. At this point probabilities in the model are transferred between unrevealed condition states and known condition states according to Equation (6). *Q(t)* and *Q'(t)* are the state probabilities immediately prior to following inspection respectively and *k* represents the states where the component is scheduled for a certain type of repair and *i* represents the state where the corresponding type of repair is necessary.

$$\begin{aligned} Q'_{\mathrm{k}}(t) &= Q_{\mathrm{k}}(t) + Q_{\mathrm{i}}(t) \\ Q'_{\mathrm{i}}(t) &= 0 \end{aligned} \quad \textbf{(6)}$$

4.5 Expected maintenance costs

Average repair costs for each type of maintenance work on each of the bridge elements of different materials were estimated from the database of previous work carried out. Note that for a certain type of maintenance, maintenance costs will include the cost of possession if necessary. The total repair cost over the structure life period is then calculated by taking the product of the number of bridge element repairs of each severity and the average costs of such repairs. The number of bridge element repairs can be calculated by integrating the rate of transitions from each corresponding degraded state to the new state over the prediction time, T. The expected repair costs are given

in Equation (7). The servicing and inspection cost are also considered, depending on the frequency of the inspections and services, these costs can easily be added to the total costs. In total, the expected maintenance cost for a component is:

$$\begin{gathered}\text{Total expected maintenance cost = Minor repair cost + Major repair cost +} \\ \text{Replacement cost + Servicing cost + Inspection cost} \\ = \int_0^T \mathrm{Q}_k^i(t).\nu_1^i dt \times C_1^i + \int_0^T \mathrm{Q}_l^i(t).\nu_2^i dt \times C_2^i + \int_0^T \mathrm{Q}_m^i(t).\nu_3^i dt \times C_3^i + \\ \sum_{f=1}^{T/f}[S_f] + \sum_{f=1}^{T/f}[I_f]\end{gathered} \quad \textbf{(7)}$$

where
T = Length of the prediction period (year)
$\mathrm{Q}_k^i(t)$ = Probability of the component *i* requires minor repair at time t and has been scheduled to repair (State *k*)
$\mathrm{Q}_l^i(t)$ = Probability of the component *i* requires major repair at time t and has been scheduled to repair (State *l*)
$\mathrm{Q}_m^i(t)$ = Probability of the component *i* requires replacement at time t and has been scheduled to be replaced (State *m*)
ν_1^i = Minor repair rates of the component *i*
ν_2^i = Major repair rates of the component *i*
ν_3^i = Replacement rates of the component *i*
C_1^i = Average Minor Repair Costs of the component *i*
C_2^i = Average Major Repair Costs of the component *i*
C_3^i = Average Replacement Costs of the component *i*
S_f = Cost of servicing after every interval of f years
I_f = Cost of inspection after every interval of f years

4.6 Model overview

Figure 6 shows an overview of the parameters that were considered in this paper to support the modelling of railway bridge asset. These parameters are taken from the database combined with maintenance decisions made by the user to form the model inputs. They go into a Markov bridge model generated to produce the state probability prediction results. A lifetime duration, over which the model predictions will be made, has been assumed to be 60 years. The initial condition of the bridge elements was taken from the database assuming the element's condition has not changed from the recorded condition at the last inspection. The simulation and mathematical integration give the probability of being in each different model state throughout the entire asset predicted lifetime.

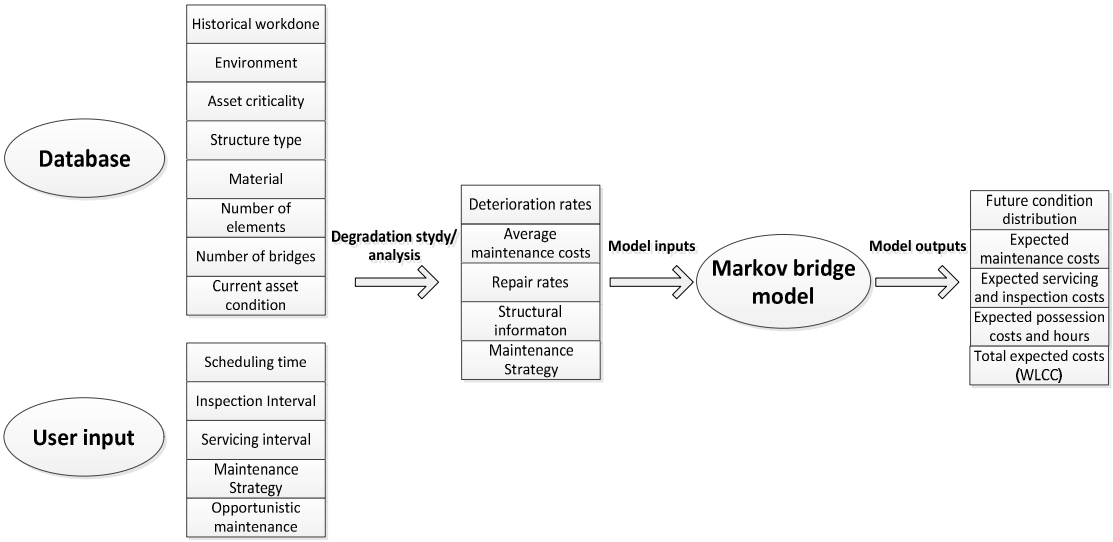


Figure 6: Complete bridge model in modelling bridge asset

5 Model implications

This section presents the results which have been produced for a chosen typical metal underbridge structure taken from the database to demonstrate the capabilities of the model. The bridge main components and their conditions are illustrated in Figure 7. A Markov bridge model of $2x4^5$=2,048 states was generated for the analysis.

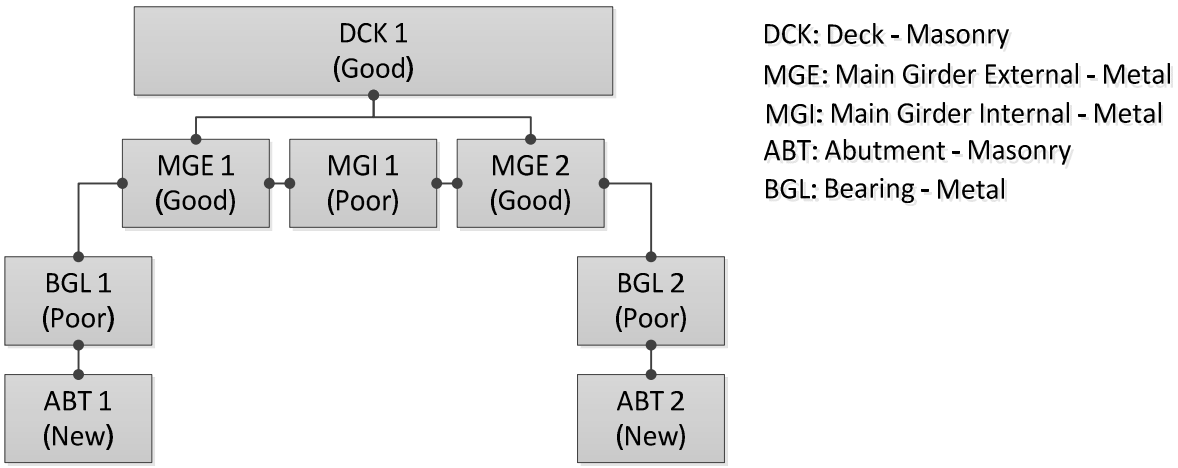
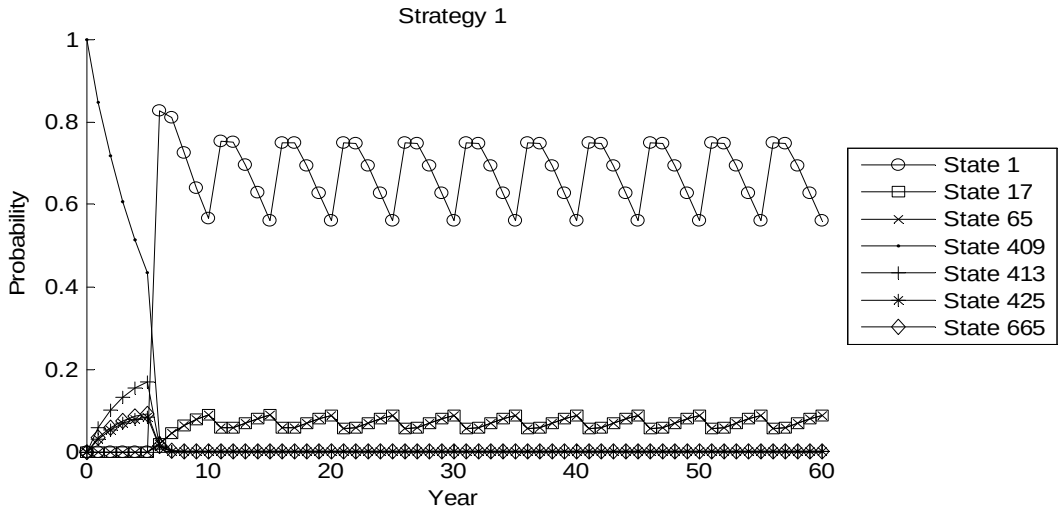


Figure 7: Structural arrangement of a metal underbridge

5.1 Strategy 1 - repair as soon as the component needs repair



States	DCK	MGE1	MGI1	MGE2	BGL1	BGL2	ABT1	ABT2
1	As New	As New	As New	As New	As New	As New	As New	As New
17	As New	Good	As New	Good	As New	As New	As New	As New
65	As New	As New	Good	As New	As New	As New	As New	As New
409	Good	Good	Poor	Good	Poor	Poor	As New	As New
413	Good	Good	Poor	Good	V. Poor	V. Poor	As New	As New
425	Good	Poor	Poor	Poor	Poor	Poor	As New	As New
665	Poor	Good	Poor	Good	Poor	Poor	As New	As New

Figure 8: Probabilities of being in different states of the bridge model under maintenance strategy 1.

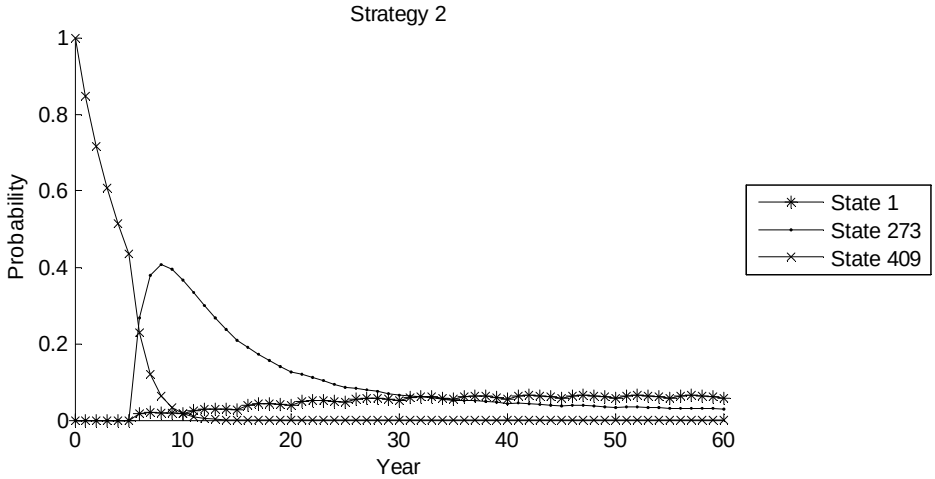
Figure 8 shows the probability of the bridge model being in different states over a 60 year period assuming maintenance strategy 1. Each state in the model is a unique combination of all the bridge element conditions, the mapping of the model state and the element conditions can be seen in the table below the graph. It can be seen that the model starts with the probability of 1 of being in state 409, this is the initial condition of the bridge elements. As the inspection period was set to every 6 years and the repair process happens after the first inspection. The figure shows that during the first 6 years, the probability of the bridge model being in states 409 decreases and the probabilities of being in state 413, 425 and 665 increases. These states represent the deterioration of the bearings, the girders and the deck respectively. By the end of the first 6 years, the probability of all the components remaining in the same condition as the initial condition is about 45%. The probability that the bearings (BGL) deteriorate to a very poor state (state 413) is almost 20% whilst the likelihood of the deck or any main girder deteriorating to a worse state is about 10% and 20% respectively. Note that the probability of being in any other states that is less than 1% was not included in the plots. As the strategy is to carry out repair as soon as the components are revealed to be in the state where any type of repair is necessary, the repair process can be clearly seen after the 6th year with the

probability of the components being in the as new condition increases and the probabilities of being in any worse condition decrease. With this maintenance strategy, there is an average probability of 65% that all the components will be operating under the as new condition (state 1).

The effects of maintenance can be seen by the ‘wave’ nature of the plot. The peak of the ‘wave’ is when the inspection happens and the condition of the component is revealed. Following this point, any revealed failures are being scheduled for repair and thus the probability of being in an ‘as new’ condition increases. A certain time after the repair, as the component continues to deteriorate, the probabilities of being in poorer condition states increases. This process is what creates the ‘wave’ shape in the plot. After the next inspection when the component condition is revealed, the process is repeated again.

5.2 Strategy 2 with and without opportunistic maintenance

Figure 9 and Figure 10 show the effects of opportunistic maintenance on repair strategy 2. It can be seen that after the first inspection, the normal maintenance strategy carried out repair on the internal main girder (MGI) and the bearings (BGL) as these components are in the condition where major repair is necessary. This process brings the system to state 273. Opportunistic maintenance not only carries out repair on components needing a major repair but also on other degraded state components, this would bring the system to state 1 where all the components are in the as good as new condition. This is reflected in Figure 10 as the probability of being in state 1 increases after the first inspection. As the result of applying opportunistic maintenance, it is more likely that the component will be operating in a better condition comparing with the case where opportunistic maintenance is not employed.



States	DCK	MGE1	MGI1	MGE2	BGL1	BGL2	ABT1	ABT2
1	As New	As New	As New	As New	As New	As New	As New	As New

273	Good	Good	As New	Good	As New	As New	As New	As New
409	Good	Good	Poor	Good	Poor	Poor	As New	As New

Figure 9: Probabilities of being in different states of the bridge model under maintenance strategy 2 without opportunistic maintenance

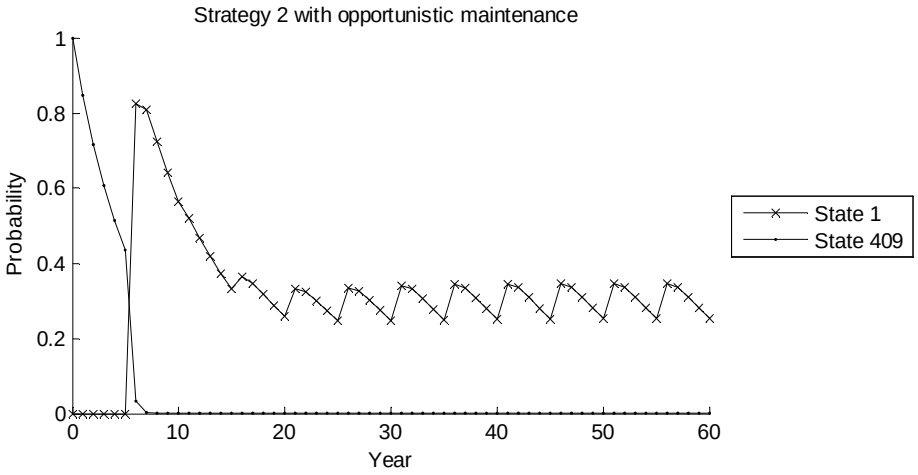


Figure 10: Probabilities of being in different states of the bridge model under maintenance strategy 2 with opportunistic maintenance

5.3 Analysis on a particular element

As well as predicting the probability of the whole bridge system being in different states, Figure 11 shows the probability plot for a particular element, metal deck, being in different conditions. The plot shows that under maintenance strategy 3, the probability of the deck being in the poor condition is about 50% after 30 years. Carrying out a similar analysis on other elements, the probabilities of the components being in a certain condition state can then be compared. This information is useful to identify critical components in the structure as well supporting the maintenance decision making process.

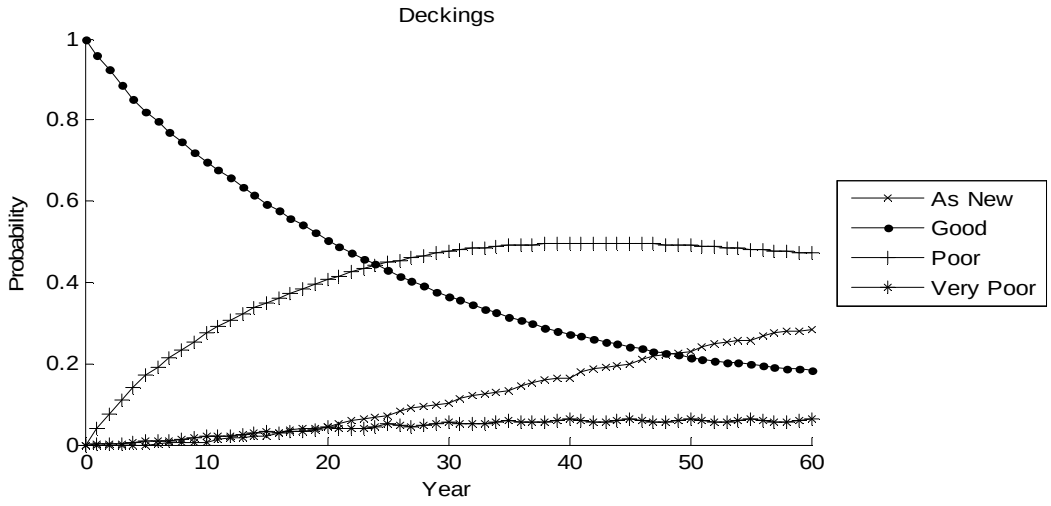


Figure 11: Probabilities of being in different states of metal deckings under maintenance strategy 3.

5.4 Expected total maintenance cost

Figure 12 shows the expected cumulative maintenance cost for all maintenance strategies. It is clear to see that following strategy 1, because the components such as deck, girders and bearing are all initially in the state where the repair is necessary hence they are scheduled to be repaired immediately after the first inspection, this results in a very high initial maintenance cost. In contrast, strategy 3 does 'minimum' work by allowing the component to deteriorate to a very poor state before intervention, the total expected maintenance cost for this strategy after 60 years is around £71k, which is almost two thirds of what is expected from maintenance strategy 1. It is worth noting that strategy 2 with opportunistic repair results in a similar initial cost as strategy 1 since all the bridge elements are scheduled for repair at the same time. In general, opportunistic maintenance results in a higher maintenance costs however the probability of an asset being in better condition is higher. Depending on a particular asset, these strategies can then be applied where the trade-off between the total expected maintenance costs and the condition profiles can be explored, allowing the most appropriate maintenance strategy to be selected.

The model is capable of investigating the effects of different maintenance strategies on the bridge. Furthermore, other model parameters such as: inspection, servicing interval, repair scheduling times can also be varied to allow a wide variety of maintenance scenarios to be explored.

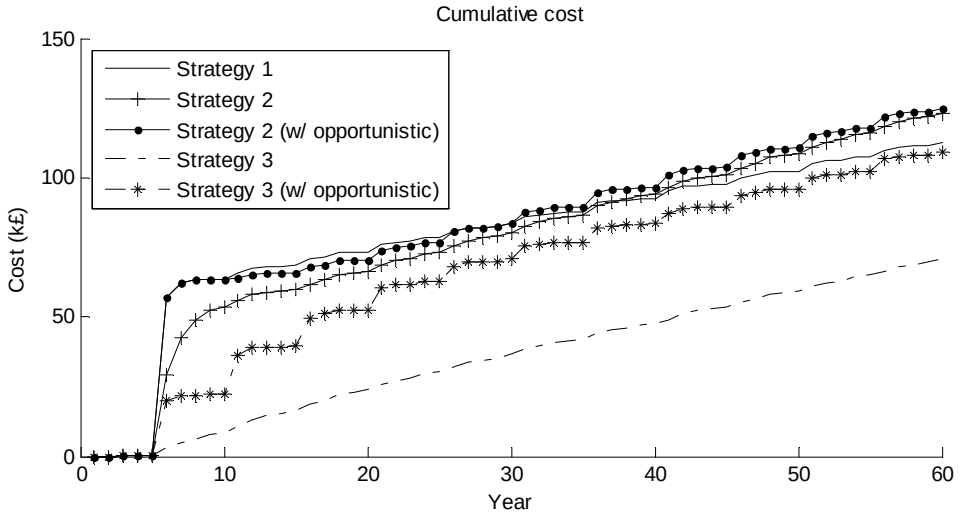


Figure 12: Cumulative expected maintenance cost for all repair strategies

6 Conclusions

This paper demonstrates a Markov modelling approach to predict the condition of individual bridge elements along with the effects that interventions such as servicing, repair, and replacement will produce. For each bridge

element the degradation process is determined by examining the maintenance records and analysing the times that each element takes to deteriorate to the point where a certain type of intervention is required. A bridge model was constructed to investigate different maintenance strategies.

The model is capable of modelling the elements accounting for: current condition, material types, structure type, asset criticality, environment, inspection intervals, servicing intervals, repair strategy and the repair scheduling (delay) times. The model outputs are the probabilities of the bridge as well as a bridge element being in different states at any given time in the future; the expected maintenance cost for each type of intervention for each bridge component; and the total expected maintenance expenditure – WLCC over the entire prediction period.

7 Acknowledgement

John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is also Director of The Lloyd's Register Foundation (LRF)[1] Centre for Risk and Reliability Engineering at the University of Nottingham. Bryant Le is conducting a research project funded by Network Rail. They gratefully acknowledge the support of these organisations.

[1] Lloyd's Register Foundation supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

8 References

1. Jiang, Y. and K.C. Sinha, *Bridge service life prediction model using the Markov chain.* Transportation Research Record, 1989(1223).

2. Robelin, C.-A. and S.M. Madanat, *History-Dependent Bridge Deck Mainenance and Replacement Optimization with Markov Decision Process.* Journal of Infrastructure Systems, 2007. **13**(3): p. 195-201.

3. Cesare, M.A., et al., *Modelling Bridge Deterioration With Markov Chains.* Journal of Transportation Engineering, 1991. **118**(6): p. 2237.

4. Ortiz-García, J.J., S.B. Costello, and M.S. Snaith3, *Derivation of Transition Probability Matrices for Pavement Deterioration Modeling.* Journal of Transportation Engineering, 2006. **132**(2): p. 141-161.

5. Chase, S.B. and L. Gaspa, *Modelling the reduction in load capacity of Highway Bridges with age.* Journal of Bridge Engineering, 2000. **5**(4).

6. Morcous, G., *Performance Prediction of Bridge Deck Systems Using Markov Chains.* Journal of Performance of Constructed Facilities, 2006. **Vol. 20**(No. 2): p. 146–155.

7. S-K, N. and F. Moses, *Prediction of Bridge Service Life Using Time-dependent Reliability Analysis*, in *Bridge Management 3: Inspection, Maintenance, Assessment and Repair*1996, E&FNSpon.

8. Kleiner, Y., *Scheduling inspection and renewal of large infrastructure assets.* Journal of Infrastructure Systems, 2001. **7**(4): p. 136-143.

9. Sobanjo, J., P. Mtenga, and M. Rambo-Rodenberry, *Reliability-Based Modelling of Bridge Deterioration.* Journal of Bridge Engineering, 2010.

10. Mishalani, R.G. and S.M. Madanate, *Computational of infrastructure transition probabilities usnig stochastic duration models.* Journal of Infrastructure, 2002. **8**(4): p. 139-148.

11. Yang, Y.N., H.J. Pam, and M.M. Kumaraswamy, *Framework Development of Performance Prediction Models for Concrete Bridges.* Journal of Transportation Engineering, 2009. **135**(8): p. 545-554.

12. Agrawal, A.K., A. Kawaguchi, and Z. Chen, *Deterioration rates of typical bridge elements in New York.* Journal of Bridge Engineering, 2010. **15**: p. 419.

13. Frangopol, D.M., S.J. Kong, and S.E. Gharaibeh, *Reliability-based life-cycle management of Highway bridges.* Journal of Computing in Civil Engineering, 2001. **15**(1): p. 27-34.

Fault Tree Analysis of Polymer Electrolyte Fuel Cells to Predict Degradation Phenomenon

Michael WHITELEY, Dr Lisa BARTLETT-JACKSON, Dr Sarah DUNNETT
Department of Aeronautical and Automotive Engineering, Risk and Reliability Research Group, Loughborough University, Loughborough, UK

Abstract:
Hydrogen Fuel Cells are an electro-chemical, zero-emission energy conversion and power generation device. Their only products are heat and electrical energy, and water vapour. One of the major hurdles to the uptake of this technology is the reliability of the fuel cell system.

This hurdle can be overcome through in depth reliability analysis including Failure Mode and Effect Analysis (FMEA) and Fault Tree analysis (FTA) amongst others. Research has found that the reliability research area regarding hydrogen fuel cells is still in its infancy, and needs development. This paper looks at the current state of the art in reliability analysis regarding Polymer Electrolyte Fuel Cells (PEMFC). A recent fault tree (FT) from the literature is qualitatively analysed to ascertain its practicality in relation to PEMFC degradation analysis.

The fault tree was found to harbour certain aspects that could be improved upon. There was no FMEA undertaken to precede the FT which would have given a greater understanding of the possible failure modes in a PEMFC system and their relationships. The FT was found to be lacking dependant relationships which are apparent in a PEMFC system. The data from the literature was also analysed to check its relevance in today's fast moving PEMFC research. Conclusions are given to the way forward for future reliability evaluation of PEMFCs.

1 Introduction

Climate change issues and sustainability concerns have increased in interest and awareness in recent years, since anthropogenic activities have been found to impact considerably upon the environment [1]. The way in which manmade activities contribute to climate change is mainly due to greenhouse gas (GHG) emissions. These include, among others, carbon dioxide (CO2), methane (CH4) and nitrous oxide (N2O) that contribute to the greenhouse effect. Additionally, energy prices are set to continue to rise by alarming rates[2] which will disrupt the UK's energy system due to a rise in oil prices.

The UK emitted 549.3 Million tonnes of Carbon Dioxide equivalent ($MtCO_2e$) in 2011[3] and 122.2 $MtCO_2e$ was due to the transport industry, with 74% of this figure due to cars, taxis and busses[4]. Due to the aforementioned negative environmental impacts of emissions from fossil fuel energy sources, this figure needs to be dramatically reduced not only to meet government targets, but for the health of the biosphere.

There are some technologies that can be used as alternatives to the fossil fuel dependent transport industry and alleviate our negative impact upon the environment. Battery electric vehicles (BEV)s have increased in popularity in recent times due to

their potential to be zero emissions (when charged with renewable power sources). However they have not been very popular due to their small ranges and long recharging time requirements[5]. These negative attributes have affected their uptake with the general public customer base, and stinted their growth and commercialization.

Hydrogen fuel cells (HFC) negate the above issues as they are an electro-chemical, zero-emission energy conversion and power generation device. Their only exhaust emissions is water that is so pure, it was used by the Apollo astronauts as drinking water on the lunar missions[6]. They can be re-fuelled in a similar time to conventional Internal Combustion Engine (ICE) vehicles, and can operate to a similar range. These positive attributes have put the HFC in the limelight as an attractive alternative to the fossil fuel dependant ICE.

At present, degradation and lifetime analysis of Polymer Electrolyte Membrane Fuel Cell (PEMFC) is sparse and still undeveloped. Although overall reliability analysis of PEMFCs is lacking in development, component level degradation data is somewhat abundant in the Fuel Cell (FC) arena. There are many useful review papers[7] [8] [9] [10] that identify the possible failure modes of a PEMFC and links to the experiments that suggest an associated degradation rate for the component.

There are certain examples of research that have aimed at addressing the holistic modelling of a PEMFC system. Rama, et al. [11] constructed a Fault Tree (FT) of a PEMFC at a qualitative level containing no data analysis, which was structured in such a way as to segregate the top failure events into the main loss pathways in a fuel cell; Activation losses, Mass Transportation losses, Ohmic losses, Fuel Efficiency losses and Catastrophic cell failure. This split the overall analysis into five separate trees to represent the ways in which a FC can fail. There is no quantitative data used in this model, thus the tree can be used to gain a greater understanding of the failure characteristics of a FC.

Placca and Kouta[12] constructed a quantitative FT of a PEMFC using aggregated data from the literature to be used to predict the lifetime of a PEMFC. They split the top event of ‘degradation of the cell’ into physical components associated with a PEMFC; Membrane, Gas Diffusion Layer and Catalyst Layer. After constructing the tree, they inputted failure and degradation data from numerous sources to predict the lifetime of a cell. A potential limitation of this method of data gathering, is that experimental results seldom use the same materials, operating conditions and parameters.

FT analysis (FTA) was most recently used in relation to PEMFC by Yousfi Steiner, et al.[13] Previously, the authors had looked into using FTA for the water management issues related to a PEMFC[14], however the recent work took a more systemic level approach. The authors used FTA to model a FC stack and its auxiliary components such as air blowers and piping. However the stack tree was somewhat basic and did not split the FC down into basic events such as in the trees of Rama and Placca & Kouta.

Even though the latest work was more systemic, it overlooked the basic events that could cause a reduction in stack voltage output which is an area that needs to be developed.

Wieland, et al. [15] used Petri-net modelling techniques to try to accurately predict the lifetime of a PEMFC stack or fleet of cars, incorporating Monte Carlo simulation techniques. The model can take into account reversible events, spontaneous events and repairable items. As with the aforementioned FT work, failure and degradation data was taken from the literature to input into the model. The model presented can be quickly and easily adapted to new situations during operation, where new transitions can be freely added. However the authors state in their concluding remarks that a lot of simplifications had to be undertaken to achieve this model. A step forward would be to address the simplifications and to map operating time realistically, which would allow for an availability analysis to be undertaken.

The current level of research relating to reliability analysis and lifetime prediction is underdeveloped and requires advancement. FTA is known for being most suited to simple systems with no dependencies between failure modes and interlinked relationships. Petri-Net analysis can take dependencies into account, however it is a lot more complex than FTA and involves a lot more computing time in comparison. Even though there is a multitude of FC degradation experimental data available, it is seldom uniform and directly comparable to other data sets.

This paper looks into the most recently proposed FTA research and analyses the techniques used and methods adhered to, to achieve the said research. The goal of this paper is to understand how to advance the PEMFC durability and degradation research area through in-depth analysis of the current research available.

2. Hydrogen Fuel Cell Overview

There are five main types of HFC that have been developed over the years, these are;

- Polymer Electrolyte Membrane Fuel Cell (PEMFC)
- Alkaline Fuel Cell (AFC)
- Phosphoric Acid Fuel Cell (PAFC)
- Molten Carbonate Fuel Cell (MCFC)
- Solid Oxide Fuel Cell (SOFC).

The main way in which they are segregated is by their constituent materials for the electrolyte based upon the operating temperature. PEMFC, AFC and PAFC have relatively low operating temperatures (<200^{o}C), and can thus utilise aqueous electrolytes, whereas MCFC and SOFC operate at temperatures from 600^{o}C and 1000^{o}C respectively, and thus cannot use aqueous electrolytes due to vapour pressure.[16]

Out of the many classifications of hydrogen fuel cell, the PEMFC is commonly singled out as the most appropriate to be implemented into an automotive application. This is due to its relatively low operating temperature of around 50-80^{o}C, its ability to use air as the cathode reactant and its rapid start-up time. The PEMFC will therefore be the focus of this paper.

2.1 Operation of a PEMFC

A fuel cell is an electro-chemical energy generation device that directly uses H_2 and O_2 to create electrical and heat energy, with the only by-product of the reaction being water. The overall reaction that takes place in a PEMFC is shown in **Error! Reference source not found.**:

$$H_{2(gas)} + \frac{1}{2}O_{2(gas)} \rightarrow H_2O_{(liquid)} + \text{heat} + \text{electricity} \quad (1)$$

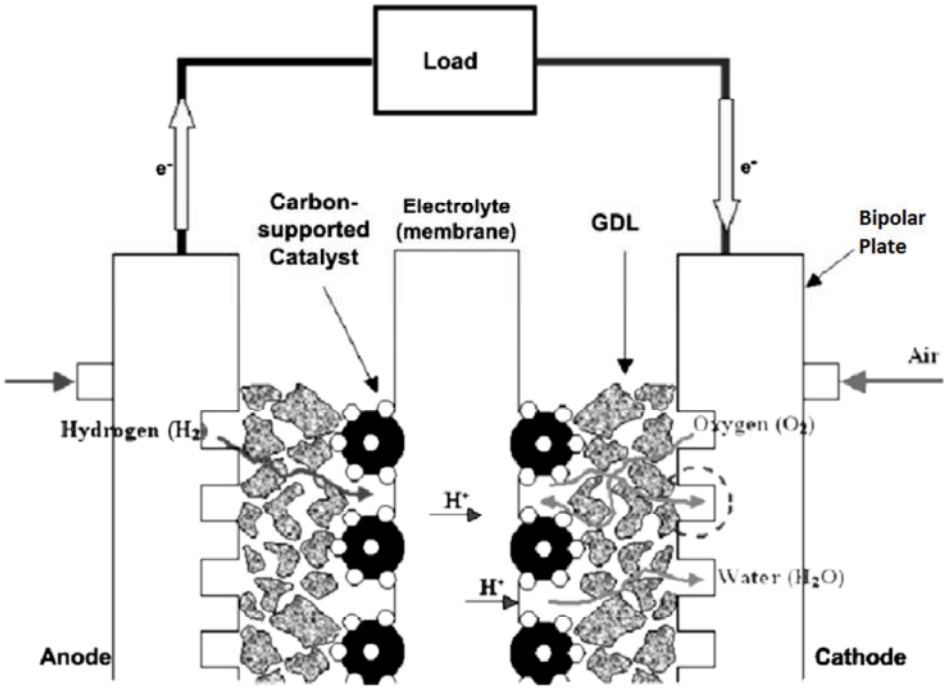


Figure 1 - PEMFC components, adapted from Kandlikar & Lu, 2009.[17]

A PEMFC is made up of four main components;

- Polymer Electrolyte Membrane
- Catalyst layer
- Gas Diffusion Layer (GDL)
- Biploar Plate

As can be seen in Figure 1, the catalyst layer, GDL layer and bipolar plates are repeated either side of the central membrane creating a 'sandwich' called a cell. The cells are then layered to create a 'stack' which forms the power plant in a PEMFC vehicle. Hydrogen gas is supplied to the anode side of the cell, and Oxygen is supplied to the cathode either in pure form, or as part of ambient air intake.

The voltage of a PEMFC can be expressed as in **Error! Reference source not found.**

$$V_{cell} = E_{nernst} + \eta_{act,a} + \eta_{act,c} + \eta_{ohmic} + \eta_{concentration} \quad (2)$$

Where E_{nernst} is the open circuit voltage potential of the fuel cell, $\eta_{act,a}$ and $\eta_{act,c}$ are the activation losses at respective electrodes, η_{ohmic} is the loss due to electrode, connections and polymer proton resistance, and $\eta_{concentration}$ relates to the losses due to concentration of fuel.

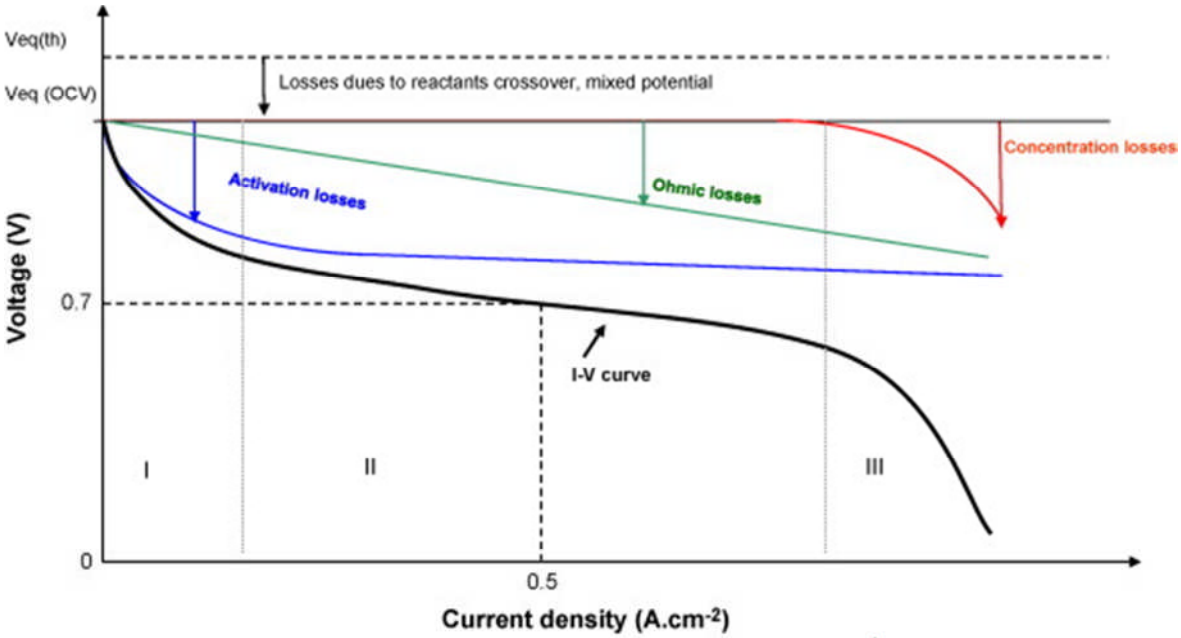


Figure 2 - Standard polarisation curve[8]

Figure 2 presents the standard polarisation curve which is used throughout FC science. It indicates the evolution of the PEMFC voltage in relation to the applied current, and the losses associated with certain fuel cell phenomenon. V_{eq}(th) shown by the dotted line at the top of the graph is associated with the fuel cell's theoretical maximum potential or Open Circuit Voltage (OCV), this is commonly known as 1.229V.

2.2 PEMFC Degradation

Currently there are three main hurdles to the commercialisation of PEM fuel cells and their competition with the ICE, these are; infrastructure, cost and durability[9]. PEMFC durability issues can be mitigated against through reliability analysis techniques. Degradation of a PEMFC is therefore a prominent area for extensive research to aid in the goal of commercialising the PEMFC car.

Degradation in a PEMFC is measureable via a reduction in output voltage.[10] Current lifetime goals accepted by the fuel cell community state that a fuel cell in an automotive application must operate for 5000 hours with a reduction in output voltage of no more than 5% over the 5000 hours period. This gives a solid indicator and timeframe for the measurement of undesirable degradation in a PEMFC.

3 PEMFC Fault Tree Analysis

3.1 Current state of analysis

Fault trees have been used by some researchers to try to gain a greater understanding of the failure modes of PEMFC's. The following sections look at a recent paper regarding a FT analysis of a PEMFC, commenting on possible

improvements and future work to develop this field further. The following work was chosen to be analysed due to the fact that it is the most up-to-date FT of a PEMFC. Previous attempts had been developed at a qualitative level[11] which have been built upon since, and as such, the latest work is chosen to be evaluated and developed further.

Placca and Kouta[12] recently used FT analysis to investigate PEM fuel cell degradation. They constructed a FT of a single cell PEMFC with the overall concerning factor being the top event of 'Degradation of the cell' to an extent that was detrimental to the functioning of the PEMFC system. They used reliability data compiled from many sources within the literature, and extrapolated the data to acquire a formatted degradation measure. From various literary sources, they came to the conclusion that there are 37 individual basic events to be considered when analysing the degradation of a PEMFC. The FT presented is a 'physical' analysis of a single cell PEMFC, splitting the top-event of the 'Degradation of the Cell' down through an OR gate into three physical components of a PEMFC:

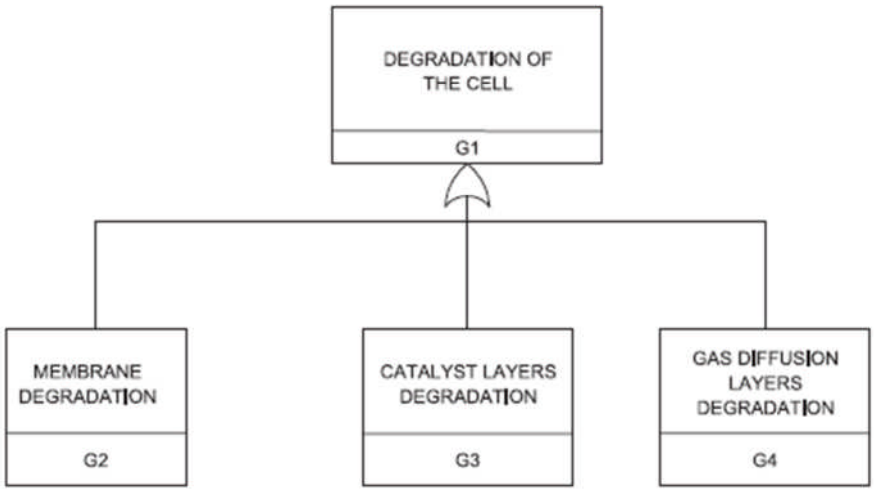


Figure 3 - 'Global' Fault Tree presented by Placca and Kouta.

These are three of the four main physical components of a PEMFC with only the bipolar plate being omitted. G2, G3 and G4 each had 12, 12 and 6 intermediate events respectively, which further branched down through OR gates to the basic events. As all of the gates in the presented FT were of the OR variety, the minimum cut sets are simply all order one's, representing the basic events of the tree.

3.2 Analysis review – Limitations Determined

3.2.1 Top Event: The top event for the tree developed by Placca and Kouta can be interpreted as vague; 'Degradation of the cell' does not directly inform the reader of what 'degradation' is classed as, and what drop in voltage output is considered to be degraded. A limitation with using fault tree analysis for this type of application is the fact that the events in the tree need to be binary in nature. A failure or a success of a component or process is an ideal scenario for using FT analysis, however the top event of 'Degradation of a cell' does not fit these criteria. A more prudent way of

defining the top event would be to suggest a rate of degradation over a time period that is unacceptable.

3.2.2 Bipolar Plate Omission: The omission of the bipolar plate component of a PEMFC is an issue that needs to be addressed. There are many studies in the literature that document and analyse the degradation and failures of bipolar plate materials in PEMFC. Failure modes affecting the bipolar plates include, but are not limited to; corrosion of the metal bipolar plate when in contact with the aqueous and acidic environment of the PEMFC, mechanical fatigue caused by repeated thermal cycles and silicone sealant used as a gasket on the bipolar plate can degrade and enter the membrane.

Corrosion and mechanical failure have been documented occurrences in a number of studies.[18,19,20] Y.C. Chen et al. showed in a 2012 study[21] that cell performance can be dramatically reduced through bipolar plate corrosion and the formation of passive oxides creating an oxide film. The corrosion of the plate material and the formation of the oxide film reduce the electrical conductivity of the plate. They showed that the film gradually increases in thickness with age, and as such the resistance increases with the thickness of the film. This phenomenon is only present on the cathode side where O_2 is the fuel and hence having the opportunity to form the oxide layer. At the anode side, H_2 is the fuel, and thus this issue is not apparent.

The bipolar plate can also affect other parts of the PEMFC, for example steel bipolar plates can release Fe^+ and Cu^+ ions that can have a detrimental affect elsewhere. The presence of foreign cations is included in the FT. For example in an intermediate event in the membrane branch named ‘Contamination by trace metal ion – G10’.

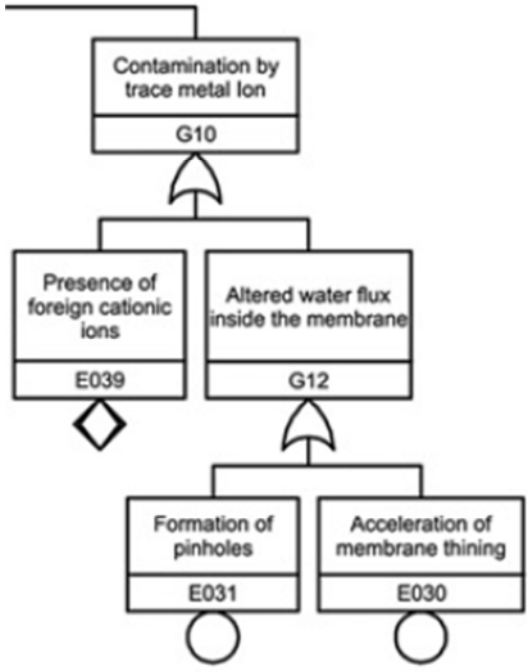


Figure 4 - Metal Ion Intermediate Event

As can be seen in Figure 4, G10 is split into the presence of foreign cations and altered water flux. The FT presented puts accelerated membrane thinning and

pinholes as the cause of altered water flux, whereas review papers[7,9] suggest that the presence of the foreign cations displace the H+ ions in the membrane, and would lead to membrane thinning and possibly pinhole formation. This indicates that the logic of this section of the membrane branch needs re-evaluating.

As mentioned earlier, foreign cations can stem from the bipolar plate material as well as inlet piping or humidifier materials. This fact, alongside the other bipolar plate corrosion issues mentioned previously, would indicate that the overall structure of the tree would need further evaluation.

3.2.3 Relationships between failure modes: One of the main limitations with current fault tree analysis of PEM fuel cells is the lack of forged links between different failure modes, where upon there exists key relationships between certain failure modes in a PEM fuel cell system.

In order to understand the links between the different basic events in the current FT, the entirety of the basic events contained in the intermediate events branches were plotted out with any potential links and relationships highlighted, the relationships are shown in Figure 5 for G2, ‘membrane degradation’.

In the figure, if one event can cause another, it is linked by a solid arrow. Additionally if one event can increase the impact or occurrence of another event, it is linked by a dotted and dashed arrow. A dashed line indicates where ‘basic events’ are the same, and as such shouldn’t be classed as different basic events. A dashed box indicates how a basic event in another branch can affect the membrane’s degradation rates.

Figure 5 shows how basic events in the membrane branch of a PEMFC FT are intrinsically linked. Some events can lead on to other events occurring, additionally some events can exacerbate other issues. To discuss this further, under the membrane degradation branch of the tree, G2 in Figure 3, gas crossover has been listed as a basic event (‘Increasing gas crossover’ – E020). H_2O_2 formation is listed as a basic event under the peroxide/radical degradation and is placed under the chemical degradation branch. In the ‘Electrocatalysts and catalyst layers degradation’ branch, a basic event of ‘Platinum dissolution’ highlights the issue of Pt nanoparticles separating from the CL and migrating to other areas of the cell. Bruijn, et al.[7] suggest that radicals are formed at either; the cathode through H_2O_2 which is formed as part of the oxygen reduction reaction, or through the decomposition of H_2O_2 at the anode through the crossover of O_2 from the cathode to anode. Additionally, they state that recent work shows how radicals can be formed through a more direct pathway as opposed to the H_2O_2 intermediary pathway, in the presence of Pt. This is where favourable conditions for degradation can be provided by a reaction between molecular H_2 and O_2 in the presence of Pt particles that have separated from the CL through electrode degradation. This shows that gas crossover, H_2O_2 formation and platinum dissolution are interlinked and therefore should not be listed as segregated basic events. Gas crossover creates H_2O_2, which causes radical introduction, and Pt nanoparticles from dissolution create radicals that attack the membrane. The radical attack causes thinning of membrane which can create pinholes and increased gas crossover. Operating conditions have been proven to exacerbate the above relationships, namely; high temperature, low humidification and high gas pressure.

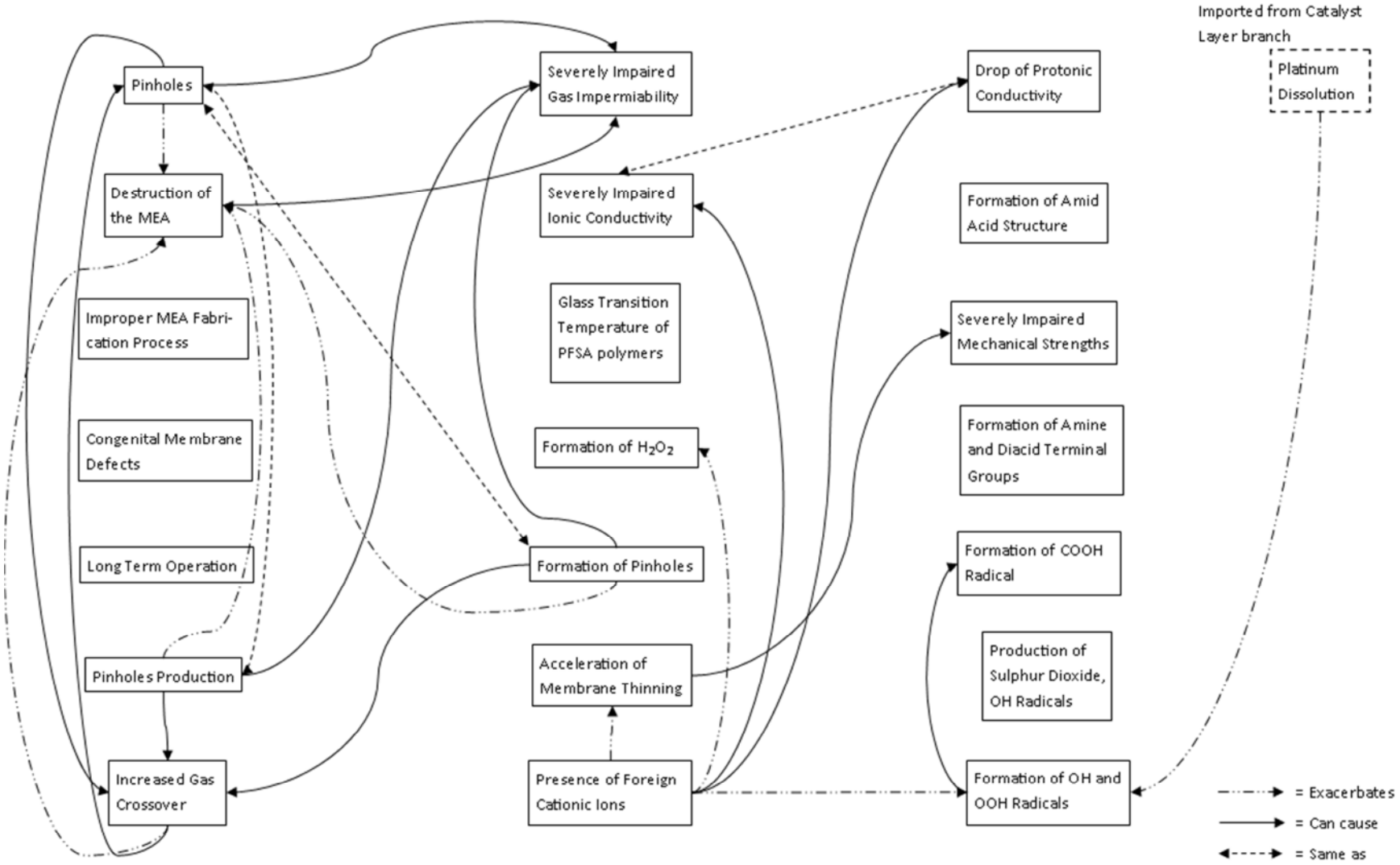


Figure 5 - Membrane Basic Event Links

3.2.4 Ambiguity of Intermediate Events: As with the top event, many of the intermediate events are not as well defined. The events in G1 – G4 are all degradation events and as such can be a small, large or complete drop of voltage. It is not clear what level of degradation is being modelled. For a more reliable FT of a single cell PEMFC, one would need to amend the current G1 – G4 events to intermediate events with binary outcomes such as 'G1 – Failure of PEMFC' and 'G3 – Failure of catalyst layer'. Degradation levels could be specified to make use of the existing failure modes, such as '60% loss of electrochemically active surface area' for the catalyst layer (G3). Basic events leading to this could include, but are not limited to: CO contamination taking up catalytic active sites, or Pt agglomeration reducing surface area of Pt catalyst for the redox reaction.

As with the previous issues found with the top events and first intermediate events, the majority of the basic events are also equivocal and are not necessarily binary in nature. E002 'long-term functioning' is ambiguous in description, not informing the reader whether it means that the cell is completely failed or is degraded to a lower output state due to long-term operation. There is no explanation of what this pertains to, such as a time frame, or what the failure mode is. During long term operation, many components can fail by any number of failure modes.

The membrane branch contains three basic events that need to be further considered. 'Pinholes', 'Pinhole Production' and 'Formation of Pinholes' are all listed as basic events, and are explained as follows; 'Pinholes' are stated as occurring 'due to exothermal combustion between H_2 and O_2'. 'Pinhole Production' is listed as not due to, but related to 'mechanical degradation'. Finally 'Formation of Pinholes' is considered to be 'due to contamination by trace metal ion'. The above would suggest that the three 'basic' events could be further broken down to fundamental basic events.

4 Possible ways of advancing the reliability study of a PEMFC

The initial FT developed by Placca and Kouta is a good first step in addressing the durability issues in PEMFC systems. However there are limitations in the study as shown by this research, which need to be addressed. The areas suggested are; Top Event, Bipolar Plate Omission, Relationships Between Events, Ambiguity of Events and Lack of Standardised Data.

4.1 Top Event

The standardly agreed criteria for PEMFC is to have a reduction in output voltage of no more than 5% over the 5000 hours period[9]. A top event reflecting this standard would alleviate the uncertainty with the current top event. It is suggested that a new top event is used to emulate the commonly accepted lifetime requirements of a PEMFC for portable applications as mentioned in section 2.2. A top event of; '5000h cell lifetime with less than 5% drop of output voltage' would clearly define the cut-off point for any values that exceed this level.

4.2 Bipolar plate omission

The bipolar plate can degrade in two main ways; corrosion which releases metal ions, and oxide film formation.

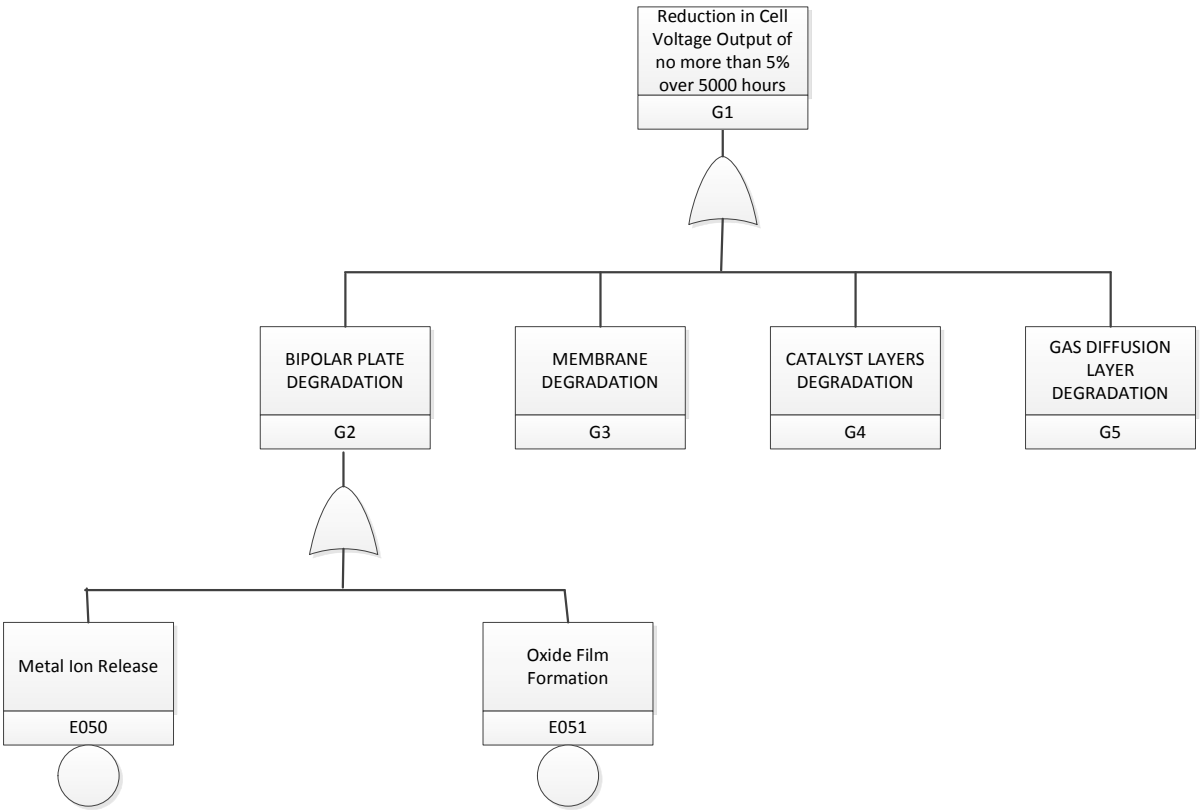


Figure 6 - Propsed change to 'global' tree

Trace metal ions such as Fe^{2+} from the bipolar plate can have adverse effects throughout the cell. It is known that these metal ions can contaminate the membrane and poison the electrode catalyst. As mentioned in section 3.2.2, the current logic regarding the trace metal ions in the membrane degradation section needs re-evaluating. The presence of foreign metal ions leads to accelerated membrane thinning and possibly pinhole generation, therefore the tree should reflect this. This would require the replication of the above basic event 'E050 – Metal Ion Release' in the newly proposed G2 and G3 intermediate events.

It has been stated that oxide film formation can increase the contact resistance of the bipolar plate by 'many orders of magnitude'[19]. This basic event would need to be included in the overall model due to its effect on the reduction of output voltage, which is the quantifier for degradation.

4.3 Relationships between events

It has been highlighted that the basic events presented by Placca and Kouta have certain relationships that make it difficult to make equivocal statements regarding the FT logic.

Pinholes were found to be three events listed as basic, but were all caused by certain conditions or phenomena. Therefore they need to be broken down further to the basic events that cause the pinholes. One way of representing pinholes in the 'mechanical degradation' branch is proposed below in Figure 7.

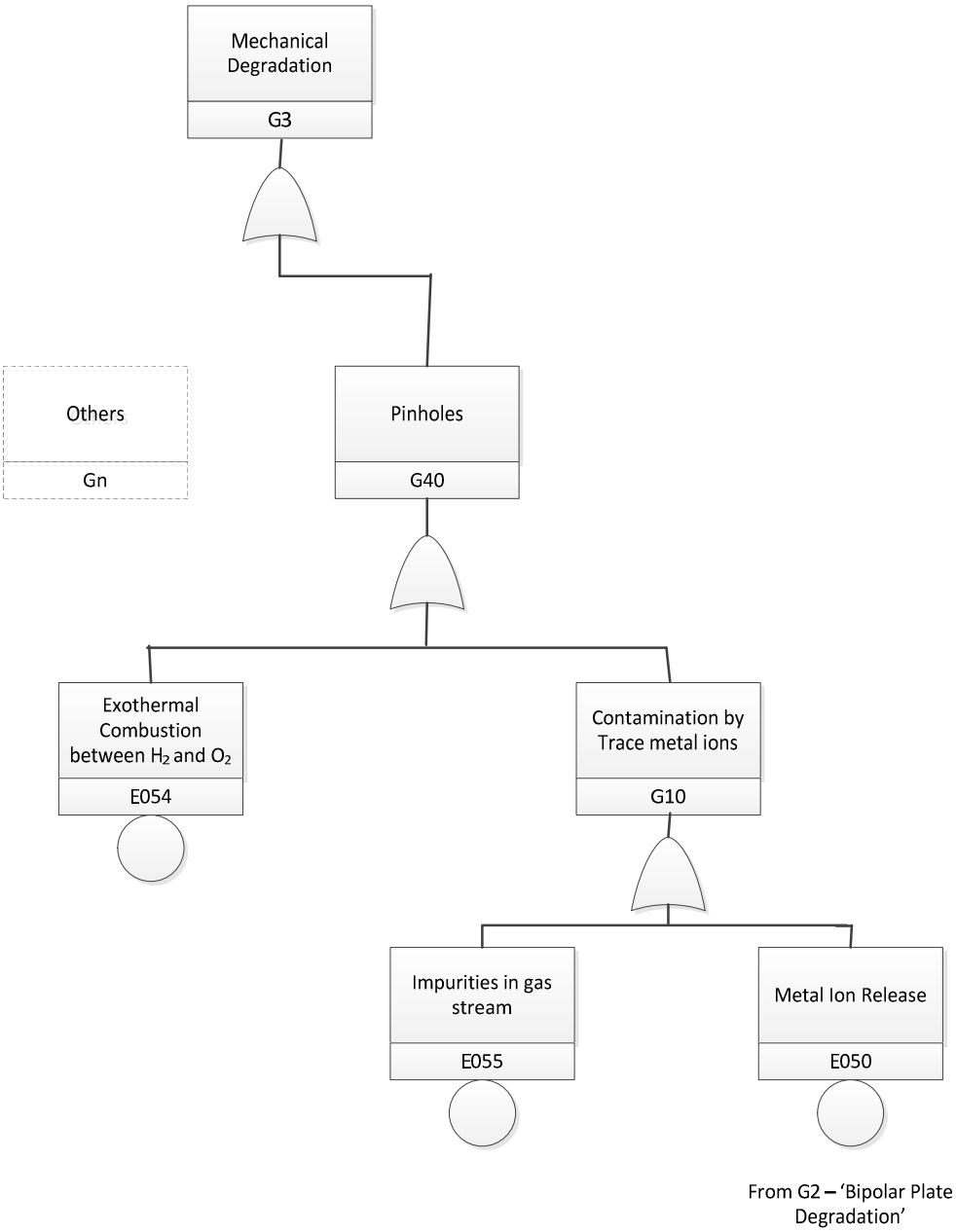


Figure 7 - Proposed pinhole logic

Figure 5 showed how certain basic events can lead on to others, and how they can also make other events worse. In particular, this instance of making others worse, questions the assumptions of the FTA techniques and its suitability for modelling these relationships with the PEMFC. It is therefore suggested that further research should look at re-evaluating the logic and structure of the presented FT. This would identify if FTA can be used for PEMFC degradation analysis, and failure forecasting.

4.4 Ambiguity of events

As with the top event, it is recommended that for a more comprehensive FT, one would need to modify all ambiguous events with definitive labels. All labels would need to be explained thoroughly or linked back to an FMEA for a clear understanding of the event.

4.5 Lack of standardised data

The lack of homogenised data for a PEMFC is a pitfall that can only be overcome by an increase in experimental analysis of certain failure modes in a fuel cell. An ideal scenario would incorporate sets of standardised experiments to homogenate degradation data, aiding with the validity of failure analysis. These would use the same cell materials, size and construction to make sure that degradation data is as reliable as possible.

5. Conclusions

Although hydrogen fuel cells have been praised as an alternative to the internal combustion engine with their 'no moving parts' slogan being used, the PEMFC of today is a highly complex machine and harbours extensively intricate relationships between components and operating conditions.

It is apparent that there is a need to aggregate the abundance of component level degradation data, into a comprehensive systemic model to forecast degradation of a PEMFC stack. This analysis will be the catalyst for a definitive way forward for modelling the lifetime prediction of a PEMFC.

A recently presented quantitative FT by Placca and Kouta has proven to be a good first step in degradation analysis and failure forecasting. Some areas that need to be addressed have been identified in this current study, in particular critical component omission, basic event logic & structure, ambiguity of events and lack of standardised data sets. It is envisaged that if these issues can be addressed, the overall degradation analysis of PEMFCs will become increasingly more accurate.

Future work will entail a full re-evaluation of the current FT logic and structure, in an attempt to account for all relationships and links. To justify the new logic and structure of the new tree, a full FMEA will be completed which will aid in the understanding of all possible links between basic and intermediate events. Close attention will be paid to potential dependencies that exist within the system whereupon a technique more suited to this analysis, such as Markov Modelling, may be incorporated within the FT.

References

[1] S. Solloman. 'Summary for Policy Makers', *in IPCC,* 2007, Viewed on 22/03/2012, http://www.lc.unsw.edu.au/onlib/refbib2.html

[2] BERR. 'DTI Energy Price Scenarios in the Oxford Models', *in National Archives,* 2010, Viewed on 22/03/2012, http://webarchive.nationalarchives.gov.uk/+/http:/www.berr.gov.uk/files/file35874.pdf

[3] DECC. '2011 UK Greenhouse Gas Emission, Provisional Figures and 2010 UK Greenhouse Gas Emissions, Final Figures by Fuel Type and End-User', *DECC,* 2012, Viewed on 31/07/2012, http://www.decc.gov.uk/assets/decc/11/stats/climate-change/4817-2011-uk-greenhouse-gas-emissions-provisional-figur.pdf

[4] DfT. 'UK Transport Greenhouse Gas Emissions', *in Department for Transport,* 2010, Viewed on 22/03/2012, http://assets.dft.gov.uk/statistics/series/energy-and-environment/climatechangefactsheets.pdf

[5] Bernardi DM, Verbrugge MW. Mathematical model of the solid-polymer-electrolyte fuel cell. J Electrochem Soc. 1992;139(9):2477-91.
[6] B. Cook. 'An Introduction to Fuel Cells and Hydrogen Technology'. *Heliocentris,* 2001.
[7] F. A. De Bruijn, V. A. T. Dam, and G. J. M. Janssen. "Review: Durability and degradation issues of PEM fuel cell components." *Fuel Cells*, vol. 8, no. 1 , pp. 3-22. 2008.
[8] N. Yousfi-Steiner, P. Moçotéguy, D. Candusso, D. Hissel, A. Hernandez, and A. Aslanides. "A review on PEM voltage degradation associated with water management: Impacts, influent factors and characterization." *J. Power Sources*, vol. 183, no. 1 , pp. 260-274. 2008.
[9] J. Wu et al. 'A review of PEM fuel cell durability: Degradation mechanisms and mitigation strategies', *Journal of Power Sources,* 184 (2008), 104-119 (pp. 105)
[10] H. Wang, H. Li & X. Yuan. 'PEM Fuel Cell Failure Mode Analysis', 1st Ed, Boca Raton, USA, Taylor & Francis Group, 2012, ch 5, sec 5.4, pp.124
[11] P. Rama, R. Chen, & J.D. Andrews. 2008. Failure analysis of Polymer Electrolyte Fuel Cells (PEFC). IN: *Proceedings of the SAE 2008*, SAE World Congress, Detroit, Michigan, April 14-17.
[12] L. Placca, R. Kouta. Fault tree analysis for PEM fuel cell degradation process modelling. *Int J Hydrogen Energy*. 2011;36(19):12393-405.
[13] N. Yousfi Steiner, D. Hissel, P. Moçotéguy, D. Candusso, D. Marra, C. Pianese, and M. Sorrentino. "Application of fault tree analysis to fuel cell diagnosis." *Fuel Cells*, vol. 12, no. 2 , pp. 302-309. 2012.
[14] N. Yousfi Steiner, D. Hissel, P. Moçtéguy, and D. Candusso. "Diagnosis of polymer electrolyte fuel cells failure modes (flooding & drying out) by neural networks modeling." *Int J Hydrogen Energy*, vol. 36, no. 4 , pp. 3067-3075. 2011.
[15] C. Wieland, O. Schmid, M. Meiler, A. Wachtel, and D. Linsler. "Reliability computing of polymer-electrolyte-membrane fuel cell stacks through Petri nets." *J. Power Sources*, vol. 190, no. 1 , pp. 34-39. 2009.
[16] EG&G Technical Services, Inc. Fuel Cell Handbook, 7th Ed. U.S. Department of Energy. 2004
[17] S. G. Kandlikar and Z. Lu. "Thermal management issues in a PEMFC stack – A brief review of current status." *Appl. Therm. Eng.*, vol. 29, no. 7 , pp. 1276-1280, 5. 2009.
[18] J. André, L. Antoni, and J. Petit. "Corrosion resistance of stainless steel bipolar plates in a PEFC environment: A comprehensive study." *Int J Hydrogen Energy*, vol. 35, no. 8 , pp. 3684-3697, 4. 2010.
[19] S. Hong and K. S. Weil. "Niobium-clad 304L stainless steel PEMFC bipolar plate material: Tensile and bend properties." *J. Power Sources*, vol. 168, no. 2 , pp. 408-417, 6/1. 2007.
[20] K. Eom, E. Cho, S. -. Nam, T. -. Lim, J. H. Jang, H. -. Kim, B. K. Hong, and Y. C. Yang. "Degradation behavior of a polymer electrolyte membrane fuel cell employing metallic bipolar plates under reverse current condition." *Electrochim. Acta*, vol. 78, pp. 324-330. 2012.
[21] Y. -. Chen, K. -. Hou, C. -. Lin, C. -. Bai, N. -. Pu, and M. -. Ger. "A synchronous investigation of the degradation of metallic bipolar plates in real and simulated environments of polymer electrolyte membrane fuel cells." *J. Power Sources*, vol. 197, pp. 161-167. 2012.

Maintenance Planning in a Saudi Arabian Hospital

Hesham Alzaben, Nottingham Trent University and Riyadh Military Hospital, Saudi Arabia
Chris McCollin, Nottingham Trent University
Lai Eugene, Nottingham Trent University

Abstract

Any organization needs maintenance activities on a regular basis to ensure safe and effective operations of its essential facilities. This is particularly true for a hospital where the primary concern is to provide a safe healthcare environment for patients. However, due to the high-risk nature of the operations involved, the maintenance department of a hospital has to develop means of ensuring equipment are operational as any unplanned interruption could adversely affect patient life. This means that potential risks contributing to equipment failure have to be identified and reduced or eliminated. Maintenance activities consume resources, it is therefore important to evaluate different maintenance approaches so that a solution that is fit-for-the-purpose can be implemented to meet specific key performance targets. Several research papers have discussed how optimization of maintenance methodologies may help to improve maintenance organization and reduce costs, thus ensuring long-term sustainability. Three important maintenance management concepts have been identified as suitable options to aid healthcare facility management. These options are Total Preventive Maintenance (TPM), Reliability Centred Maintenance (RCM) and Reliability Centred Failure Analysis (RCFA).

The aim of this project is to understand past and current maintenance-related issues, to assess how they impact on business operations, and to identify a means to improve maintenance operations in Saudi hospitals through the development of a new framework incorporating aspects of the three concepts mentioned. Review of applications of engineering management theory and practice and their mapping from the industry sector to the healthcare sector have highlighted a number of key issues which require detailed consideration, namely management policies, working environment, social culture, working culture and decision support systems among others. It became apparent that the framework will need to include a different maintenance performance measurement structure appropriate for hospital environment using approaches including high and low level performance indicators, Six Sigma and Theory of Constraints (TOC).

1.Introduction

Healthcare is the fastest growing service in both developed and developing countries. With the explosive development of knowledge, technology and globalization there is now an increasing requirement of high-technology medical

care. Every country is striving hard to cope with this increasing need of healthcare facilities in terms of both human and material resources (Feeney and Zairi, 1996). One of the most important needs for a community is to provide an adequate health care system (Haq, Hafez, 2009), which has to meet people's expectations and requirements. The success of a healthcare system can affect the reputation of a government and most governments particularly those in the developed nations have allocated substantial amount of resources to improve their healthcare systems. As a developing country, the Saudi government has over the past decade invested around US$140 billion to improve the healthcare provisions for her citizens (Saudi gazette, 2011). In 2010 and 2011 there was a substantial increase in the healthcare budget, which increased from SAR 30 billion (6.3% of total Government Budget) in 2008 to SAR 52 billion in 2009 (11% of total Government Budget) and to SAR 61.2 billion in 2010 (11.3% of total Government Budget). The budget allocation was further increased to SAR 68.7 billion, SAR68.5 billion, SAR100 in 2011, 2012 and 2013 respectively (spa.gov.sa, 2013). Despite the huge investment, there is still a shortage of healthcare services that could be provided. Moreover, the strategic direction and policies advocated by the Ministry of Health have been criticized by the Shura Council. For example, the Ministry rents 81% of buildings compare to 19% ownership, and maintenance of rented properties is often more expensive due to the restrictions imposed by the property owners (Aleqt, 2011). The Government, through the Ministry of Health (MOH), and a number of semi-public organizations who specially operate hospitals and medical services for their employees, primarily manages the Healthcare sector in the Kingdom of Saudi Arabia.

In term of improving the patient care, ministry of health has announced a new vision, which is called "patient first" to reflect its efforts and governments supports. Moreover, 24 Hospitals opened last year and there are 19 hospitals and five new medical cities under construction (MOH, 2012). With regard to operational budgeting, the Ministry of Health has assigned a huge budget to run the new hospitals and provide the highly skilled employees and high technology of equipment as well as safety and environmental facilities. According to Boussabaine et al. (2012), operation and maintenance cost of healthcare facilities (without cleaning) represents about five per cent of hospital budget in France. Therefore, to ensure the provision of high standard of hospital operation, minimization of these expenses is extremely significant as especially non-add value costs to the patient. Most hospital managers reduce the maintenance budget as a first action when they face a financial crisis. Random cutting of these budgets will lead to the reduction of the efficiency of the service provided to patients.

In this study, an attempt to evaluate maintenance management in Saudi Arabia Healthcare and use Riyadh Military Hospital (RMH) as a case study to reevaluate the facility department and establish new framework that is able to improve maintenance works. The aim of this framework is to design quality and maintenance service measurement, identify the problems which cause faults'

accumulation and repetition, create new maintenance methodology which was successfully implemented in manufacturing industry to be implemented in healthcare, study the cooperation effect between facility management departments and how to improve it to achieve work improvement and identify maintenance objectives which lead to integration with organization strategy.

2. Literature review/background research

Only one third of organizations seriously consider good maintenance management practices and realize the full benefits (Salonen and Deleryd (2011). Effective maintenance management is a growing concern of the UK manufacturing industry. Research has shown that the number of companies using a proactive maintenance approach in the UK manufacturing business has increased significantly over the last few years. Their study has shown that that the maintenance approach and continuous improvement are highly associated with effective maintenance management. 40% of the organizations participating in a pilot study still do not realize the importance of effective maintenance, management (Cholasuke, 2004). Maggard and Rhyne (1992) point out that the maintenance can represent between 10 and 40 percent of the production cost in a company. Ahlmann (2002) has estimated that the total cost of maintenance in Sweden constitutes 6.2 percent of the industry's turnover, which is close to 200 billion SEK per year. Moreover, Wireman (1990) claims that as much as one third of the maintenance cost is unnecessarily spent due to bad planning, overtime costs, bad use of preventive maintenance, etc. Failures in production systems may cause high losses, for instance in the form of lost production time or volume, negative impact on the environment, lost customers, warranty payments, etc. (Todinov, 2006).

Due to the rapid improvement of technology and complexity of machines, many organizations have been re-engineering their perspectives toward operation management. They noticed that maintenance jobs have a strong relation with profit and service interruptions. Service breakdowns not only affect on their growth but also on budgeting and customer satisfaction. Garg and Deshmukh (2006) point out that next to the energy costs, maintenance costs can be the largest part of any operational budget. Sherwin (2000) reviews maintenance organization models TPM, RCM and Total Quality Management (TQM) etc. and suggests maintenance can be a contributor to profits by use of information technology (IT).

According to Graban (2012), hospitals worldwide face a wide range of problems and pressures that have inspired them to look outside of healthcare for solution. Payers, ranging from government agencies to private insurers, are forcing price reductions on hospitals, which require hospitals to reduce costs in order to maintain their margins. Even not-for-profit hospitals need to have a surplus to remain financially viable and to drive future growth. Hospitals are becoming less

able to demand ‘cost plus’ pricing that pays them for their efforts as opposed to being paid flat rates based on patient diagnoses.”

Quality in healthcare is usually assessed by three parameters, namely, structure, process and outcome of healthcare services (Donabedian, 1988). Gelnay (2002) considers healthcare facility management (FM) as one of the key elements for the successful delivery of healthcare services. Nesje (2002) examined the distribution of FM expenditures at St Olavs Hospital in Norway, and found that maintenance, energy and cleaning costs each account for one third of the total operation costs of the hospital. In hospitals, different building systems and components, such as medical gases, fire protection systems, electricity, etc. must exhibit high levels of performance, since any minor breakdown may lead to both casualties and financial losses (Shohet and Lavy, 2004).

According to Vasseur and Llory, (1999), there are five maintenance policies covering all types of maintenance efforts: use-based maintenance (UBM), condition-based maintenance (CBM), failure-based maintenance (FBM), design-out maintenance and detection-based maintenance. However, maintenance concepts vary from organization to organization. A maintenance concept generally consists of a process and a framework, which are the supporting structures needed to manage the maintenance function in an organization (Marquez and Gupta, 2006). To develop a maintenance concept, the company must first review its operation (Marquez and Gupta, 2006). It is important to note that maintenance policy selection is based upon expert judgment on risk and related to a functional failure (Rosqvist et al., 2009).

Implementing of a maintenance framework is not an easy mission as it needs to involve the top management as well as the whole related department. Therefore, Naughton and Tiernan, (2012) point out eight difficulties in framework and individual conceptualization as the following: Cost of the development process, Lack of management support , Fear of change, Lack of plant/process-specific knowledge, Lack of respect for knowledge-based experience, Experience-based protectionism and Low morale and skepticism, Lack of respect for the practitioner. Naughton and Tiernan, (2012) outline a proposed nine-step framework for developing/ implementing an individualized maintenance management strategy. 1) Focus on the positives and define your position (2) Identify constraints and limitations (assessing complexities)(3) System classification (4) Machine classification (5) Policy selection (6) Align performance indicators (7) Structure maintenance data (8) Implement and monitor (9) Feedback. Crespo Márquez et.al’s (2009) introduced generic model maintenance management framework. This framework consists of eight sequential management building blocks.

According to Mosadeghrad et al (2008), many hospital managers have little understanding of how to satisfy their employees and how these employees’ satisfaction levels influence their intent to leave their positions. Therefore, he suggest that management and supervision, recognition and promotion, job

security and task requirement are the best predictors of job satisfaction among hospital employees, whereas the highest dissatisfaction level occurs in the area of working conditions, salaries and benefits, recognition and job security. They revealed a positive relationship between employees' job satisfaction and organizational commitment

Codinhto et al, (2008) refer to maintenance and quality of equipment as a service of hospital, which provide a stable environment condition. They determined that healthcare outcome could be affected by more than one built environment characteristic such as stress level affected by temperature. Whereas, one built environment characteristic may affect several healthcare outcomes such as light affecting depression, melanoma and retinopathy. Alturki (2011) states that major stakeholders and management commitment is essential for the successful development of a maintenance strategic plan. As maintenance activities are known as intangible benefits to an organization, any efforts, which need to develop, may meet with some stakeholders' reluctance.

3. Current State of The Hospital

In Saudi Arabia, during the construction boom, new cities were constructed and new projects were handed over. Therefore the demands on maintenance have increased recently and many organizations have established maintenance and operation departments. Additionally, maintenance management focuses on reviewed, analyzed and recommended ways of increasing the effectiveness of their maintenance management system.

Riyadh Military Hospital is one of the organizations that noticed that the maintenance department is a significant department required to maintain the hospital's buildings and equipment. This maintenance is a significant step to achieving the hospital's targets of providing an excellent service to its patients.

The maintenance department in Riyadh Military Hospital is a section of the technical affairs department, which is called the Facility Department. This department is responsible for maintenance activities in three different buildings, which are called main hospital, South West Corner Hospital (Maternity Hospital) and VIP Hospital. Moreover, it communicates with other departments in the process of ordering any new equipment and is also responsible for any alteration work around the hospital.

The Facility Department consists of the Head of Department and his secretary, Finance Officer and three Hospital Engineers with their assistants and secretaries. The Facility Department supervises an operation and maintenance company that is responsible for providing highly skilled staff to operate and maintain the hospital's facilities. The operation and Maintenance Company have a project manager with his administration team and four engineers; mechanical, electrical, civil and equipment. For each engineer there are different supervisors

with technicians and helpers. The total number of company staff is approximately 400.

When the Maintenance Department receives a complaint; the docket is issued to the supervisor who is responsible for the complaint's area. The maintenance team is then sent to the failure's location and either tries to return the equipment to its normal condition by fixing the fault or reports back to their supervisor of their inability to fix the fault. The supervisor tries to solve the problem and gives advice to his team. However, if the fault is not fixed, either due to the unavailability of the spare part or to the team's lack of skill in discovering the fault, the supervisor asks the Hospital Engineer to provide the spare parts. The Hospital Engineer is required to contact the equipment agent if the spare parts are not found or if the fault was still not resolved. Furthermore, if the repair cost was expensive or the fault was still not resolved, a rejection form is issued and a replacement order is filed to the Head of Department.

4. Research Methodology

In order to obtain the necessary information to answer the aims of the study, both primary data and secondary data research was conducted. Secondary data about the healthcare situation in Saudi Arabia and driver forces to improve that situation have been gathered throughout Saudi Arabia ministry of health web site and its publications. Moreover, the RMH publications, books, journals and articles as well as its web site were a good source for gathering secondary data information about RHM. Primary research involved gathering information about maintenance management in RMH and how to improve it. Therefore, questionnaire, group meeting, interview and observation surveyed maintenance staffs. A pilot questionnaire was discussed and seven participants completed the questionnaire in order to get some idea about the existing situation in the hospital's leadership and management, organization culture and maintenance management and process.

5. Critical successful factors

To date, critical success factors (CSFs) for implementing maintenance management (MM) in healthcare industrial have not been systematically investigated. Existing studies, which have been done in healthcare industrial, was focused on quality management.

This research provides an empirical study on the identification of the critical success factors (CSFs) of maintenance management system implementation in RMH. Through a comprehensive and detailed analysis of the literature, 17 success factors were identified to develop a questionnaire as shown in table 1. These items were empirically tested by data collected from 6 employees in RMH. After distributing this list to 7 senior maintenance workers to rank the factors, table 2 illustrates their ranking.

NO.	Critical Success Factors	1	2	3	4	5	6	Mean	Rank
1	Top management support is essential for the successful development of the maintenance system and to overcome obstacles to the application and demonstration of the importance of having a high efficiency maintenance system to achieve the organization's goal in terms of making sure the delivery of health care to the patient and reduce breakdowns and costs.	3	2	1	1	9	3	3.1667	1
2	A maintenance strategy plans compatible with the organization's goals.	1	4	5	5	1	11	4.5	2
3	A well-organized structure of the organization either hierarchically or horizontal.	4	6	12	4	2	1	4.833	3
4	Clarity of policy and procedure for each process as well as job description and responsibilities.	5	5	14	2	3	2	5.1667	4
5	Having employees with high technical skills and ability to manage the maintenance work and repair the faults	9	1	2	6	5	12	5.8333	5
6	There must be a sufficient number of employees; ensuring they are working according to their qualifications and job description	7	3	3	7	6	13	6.5	6
7	A good motivation system that encourages innovation and improvement including annual increases and bonuses	10	8	8	8	4	4	7	7
8	The existence of frequent training programs for all maintenance personnel helping them to keep updating with the new technology.	8	7	13	3	7	5	7.1667	8
9	The clarity of the maintenance contract which includes number of contract employees, salaries and qualifications	2	13	7	13	11	2	8	9
10	Attention to mental well-being and staff morale	15	9	4	14	10	6	9.667	10
11	Good performance management indicators to measure the performance of staff and equipment. By implementing these measures it is able to manage and improve it.	13	10	9	10	8	8	9.667	11
12	Need to pay attention to change management and how to apply it to reduce resistance to change and this change would be in the interest of the organization and due to the requirement of current circumstances which needs improvement and development	6	11	16	12	16	9	11.667	12
13	The need to respect the efforts of maintenance work and provide support to them by Top management and heads of departments as well as physicians support.	11	16	6	9	13	16	11.833	13
14	Promoting teamwork and sharing of information and experiences.	12	14	17	15	15	7	13.333	14
15	Limit out-sourcing maintenance work and instead of that improve the skills of existing staff to achieve maintenance's goals	16	17	10	11	12	17	13.833	15
16	Good Modern information systems to improve staff communication (wired and wireless telecommunication), control devices, control and management of maintenance work as a building management system and CMMS.	14	15	11	16	14	14	14	16
17	Achieve customer satisfaction including patients, attendants and other staff such as physicians, nurses, administrators etc.	17	12	15	17	17	15	15.5	17

Table 1. Critical success factors

For table 1, top management support has the top ranking with 3.1667, then a maintenance strategy plans compatible with the organization's goals which has 4.5 and the third factors which has the highest affect on maintenance success in hospital is a well-organized structure of the organization either hierarchically or horizontal with 4.833 mean.

From the above list, main CSFs can be identified. Main CSFs refer to factors, which are really link to organizational success. Therefore, financial, customer satisfaction, motivation, contractors, processes, safety, CMMS and education and training are the main CSFs, which have to be met when maintenance activities are established.

6. Theory of Constraint:

In the case of fault accumulation and delay in work completion in a sensitive area like a hospital, the constraint would be management related. There are several reasons given why an organization would fail to reduce the equipment' breakdown. In order to increase the efficiency of hospital equipment that are high technology and patient effective, all steps must be examined together to determine the constraint; the core problem for termination. Taylor III and Poyner (2008) point out that, while the TOC was developed for manufacturing through Goldratt's Thinking Process, the Thinking Process system can be used to work for selection and focus.

6.1. The thinking process and the theory of constraints

Goldratt (1992b) developed the Thinking Process to identify the system's constraints, which limits the organization's ability to meet its goals. One of the thinking processes in theory of constraints, which was introduced by Dr. Goldratt, called the Current Reality Tree (CRT) (Scheinkopf, 1999).

According to Taylor III and Poyner (2008), by pinpointing the problem, determining a workable solution, and implementing the solution, the thinking process is able to underscore the importance of a systematic process of problem solving. Therefore this process could be able to identify the underlying cause of the problem. These causes present undesirable effects (UDE), which can be used to develop a CRT.

6.2 Undesirable effects

In order to create CRT, the first step in the thinking process is to develop a list of at least 10 –12 undesirable effects. The questionnaire and group meeting as well as informal gathering information have been reviewed to come up with list of hospital problems as the following:

UDE 1 Lack of spare parts

UDE 2 High prices of spare parts

UDE 3 Technicians do not have the ability to get the job done professionally

UDE 4 The delay in the completion of the work

UDE 5 Repeated breakdowns

UDE 6 Maintenance department provides additional services, which do belong to them.

UDE 7 Old equipment and devices

UDE 8 Lack of modern technology

UDE 9 Accumulation of faults

UDE 10 Delay in their work which affect on other departments performance

UDE11 Awarding projects to contractors who do not have the experience to complete the work as required

UDE 12 Lack of cooperation between the technicians and supervisors

UDE 13 Frequent overtime

UDE 14 No motivation to finish the work

UDE 15 No encouragement for innovation

6.3 The Current Reality Tree

In order to draw this tree, UDEs were connected by organizing the UDEs into an effect-cause-effect relationship analysis as shown in Figure1. The tree reads as follow: If manager does not respect the work done, no cooperation between technicians and supervisors and no encouragement for innovation, then lack of motivation to finish the work. And so on.

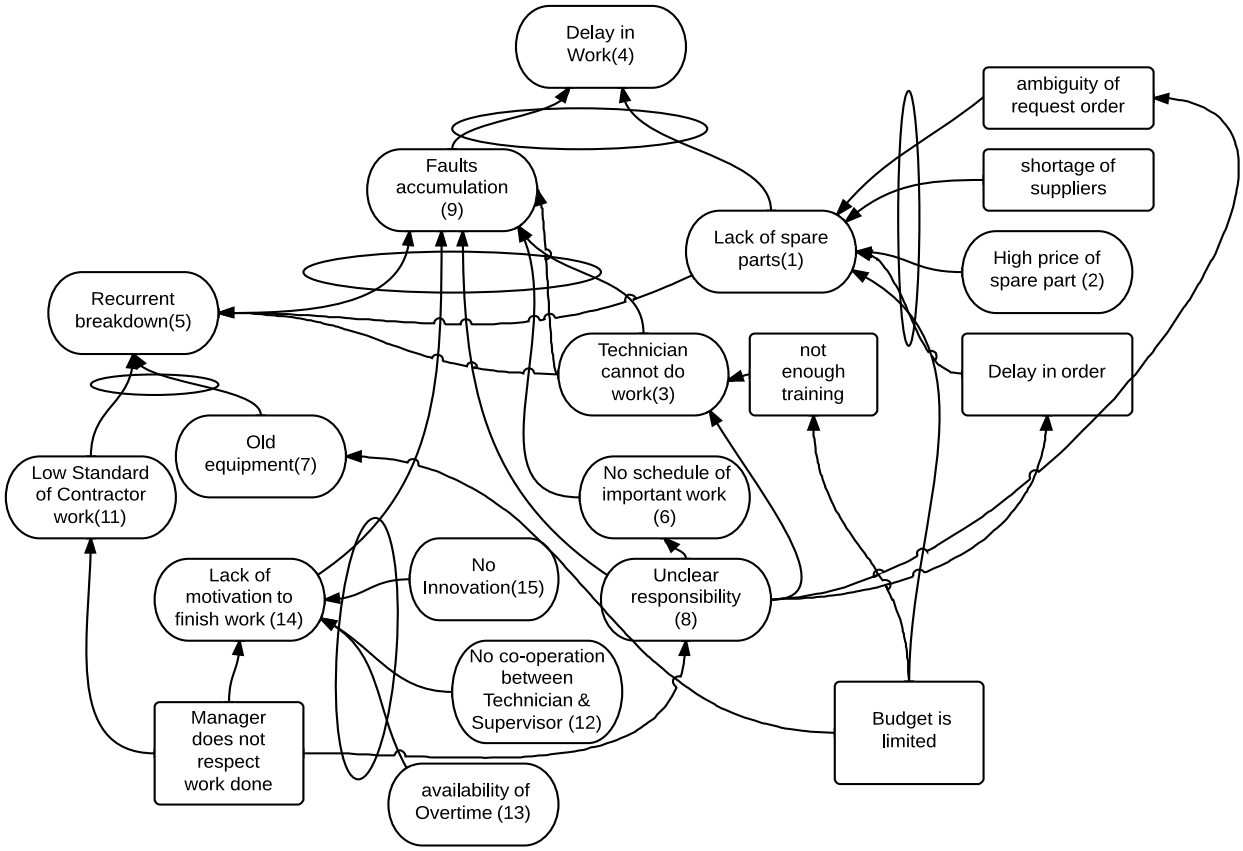


Figure 1. Current Reality Tree

In order to identify the core causes(s) of problem, “so what” test was used to distinguish between what is pertinent and what is not? Therefore, 6 entities were selected which show strong causes of the problem. The revised list of pertinent entities became:

1- Lack of spare parts
2- Technician cannot do work
3- Unclear responsibility
4- Faults accumulation
5- Low standard of Contractor work
6- Lack of motivation to finish work

Entities not involved in connecting the pertinent entities with each other need to be classified under an entry point as shown in Figure 2. Determination of the degree to which entry point is responsible for the existence of the pertinent entities was identified as shown in table 2. Scheinkopf (1999) points out that an entity point, which is responsible for 80% or more of the pertinent entities, is a core cause. Therefore, entity Manager does not respect work done (A), which is a cause for more than 80%, was selected as the core cause.

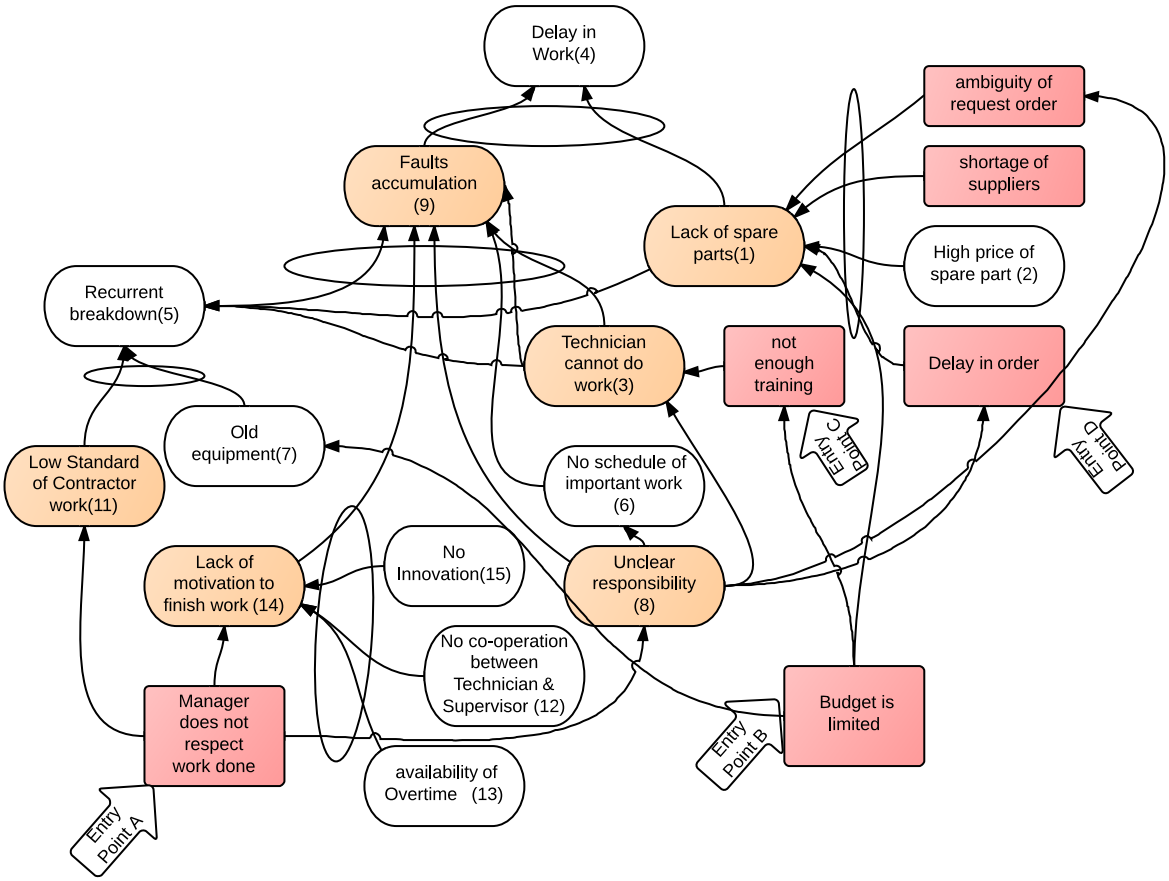


Figure 2. CRT Entity points

EP\PE	1	3	8	9	11	14	Total	%of 6
Manager does not respect work done (A)	-	X	X	X	X	X	5	83%
Budget is limited (B)	X	X	-	X	-	-	3	50%
Not enough training (C)	-	X	-	X			2	33%
Delay in order (D)	X	-	-	X	-	-	2	33%

Table 2. EP and PE relationship

7. Conclusion and future work:

Through detailed surveys, 17 Success Factors (SFs), which contributed to good maintenance practices, have been identified. By means of ranking, 9 factors, which had the support of the top hospital management, were considered critical to the success of the maintenance operations.

From collected data, a number of key issues have been analyzed by means of Current Reality Tree (CRT) and Thinking Process (TP), both of whom form part of The Theory of Constraints (TOC). The CRT analysis suggests that delays in maintenance work at the case study hospital are attributable to 3 causes - low maintenance team motivation, limited budget and low standard of contractor work. The TP analysis suggests that education/training and better communications are potential solutions to addressing some of the issues. The spare parts problem needs to be shown as details in a separate diagram, as we need to draw each problem in separate diagram as well as the whole in one diagram. The motivation problem could be initiated as a low salary, no supervision, too busy during doing jobs, no appreciation, delayed pay, no vacation and no encouragement to provide solution. To understand fault accumulation, failure data needs to be studied to recognize the fault situation. Drawing to show relationship between numbers of breakdowns, amount of time breakdown per month and deferent equipment could be helpful. Study the state of complaint, is it major and systematic, or minor and systematic or major and one off or minor and one off could be a good method to distribute the efforts of maintenance team.

An initial new maintenance framework has been formulated based on Crespo Márquez et.al's (2009) framework. The initial framework focuses on 8 primary areas as shown in figure (9.2): data collection, KPIs/PIs, strategies/policies, risk/cost activity planning, integration of maintenance practices, maintenance-related process issues, automation and new methodology.

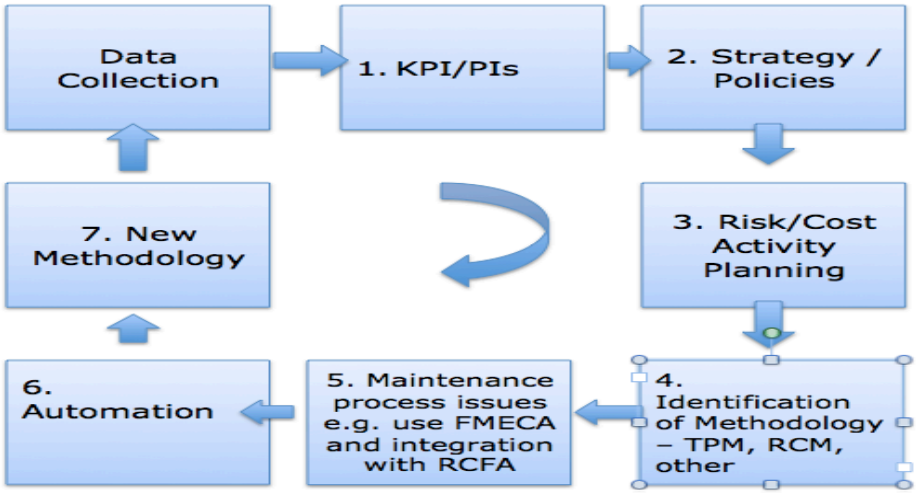


Figure 3. Proposed maintenance framework

Reference

1. Ahlmann, H.R. 2002, “From traditional practice to the new understanding: the significance of life cycle profit concept in the management of industrial enterprises”, Proceedings of the International Foundation for Research in Maintenance Management and Modeling Conference (IFRIM-2002), Vaxjo, Sweden.
2. Al Turki Umar.2011 A frame work for strategic planning in maintenance. Journal of Quality in Maintenance Engineering Vol. 17 No. 2, 2011 pp. 150-162.
3. Aleqt 2011 [One line] Available http://www.aleqt.com/2011/07/05/article_555941.html[1/7/2011].
4. Boussabaine Halim. Sliteen, Samer, Catarina, Orlando, (2012),"The impact of hospital bed use on healthcare facilities operational costs: The French perspective", Facilities, Vol. 30 Iss: 1 pp. 40 - 55
5. Codinhto et al, 2008.The impact of the building environment on health outcomes,
6. Crespo Márquez, A. P. Moreu de León, J.F. Gómez Fernández, C. Parra Márquez, M. López Campos, (2009) "The maintenance management framework: A practical view to maintenance management", Journal of Quality in Maintenance Engineering, Vol. 15 Iss: 2, pp.167 - 178
7. Donabedian, A. (1988), “The quality of care – how can it be assessed?”, Journal of American Medical Association, Vol. 260, pp. 1743-8.
8. Feeney, A. and Zairi, M. (1996), “TQM in healthcare”, Journal of General Management, Vol. 22 No. 1, pp. 35-47.
9. Gelnay, B. (2002), “Facility management and the design of Victoria Public Hospitals”, in Proceedings of the CIB Working Commission 70 – Facilities Management and Maintenance Global Symposium 2002, Glasgow,pp. 525-45.
10. Goldratt, E. (1992b), The Goal: A Process of Ongoing Improvement, 2nd ed., North River Press, Great Barrington, MA.
11. Graban, Mark. 2009.Lean Hospitals: Improving Quality, Patient Safety, and Employee Engagement.CPC Press .Newyok
12. Haq, Z.Hafeez, A. (2009) Knowledge and communication needs assessment of community health workers in a developing country: a qualitative study. Biomed Central Ltd.
13. Maggard, B. and Rhyne, D. (1992), “Total productive maintenance: a timely integration of production and maintenance”, Production and Inventory Management Journal, Vol. 33, Fourth Quarter, pp. 6-10.
14. Marquez, A.C. and Gupta, J.N.D. (2006), “Contemporary maintenance management: process, framework and supporting pillars”, The International Journal of Management Science, Vol. 34 No. 3, pp. 313-26.
15. MOH, 2012 online< lhttp://www.moh.gov.sa/EN/Pages/Default.aspx>
16. Mosadeghrad, Ali. Ferlie, Ewan. Rosenberg, Duska. 2007 A study of the relationship between job satisfaction, organizational commitment and turnover intention among hospital employees. Health Services

Management Research. pp.: 211–227. Royal Society of Medicine Press
17. Naughton, Michael Daragh, Tiernan, Peter, (2012),"Individualizing maintenance management: a proposed framework and case study", Journal of Quality in Maintenance Engineering, Vol. 18 Issue: 3 pp. 267 - 281
18. Nesje, A. (2002), “Management, operation and maintenance costs of hospital buildings”, Proceedings of the International Federation of Hospital Engineering, Bergen, pp. 290-96.
19. Narasimhan, K. (2006) "Inventive Thinking through TRIZ: A Practical Guide", The TQM Magazine, Vol. 18 Iss: 3, pp.312 - 314
20. Neely, A.D., Mills, J.F., Gregory, M.J. and Platts, K.W. (1995) ‘Performance measurement system design–a literature review and research agenda’, International Journal of Operations and Production Management, Vol. 15, No. 4, pp.80–116.
21. Saudigazette (2011) [One line] Available http://www.saudigazette.com.sa/index.cfm?method=home.regcon&contentID=2011032396529[1/7/2011].
22. spa.gov.sa, 2013 on line from http://www.spa.gov.sa/Search.php?pg=1&s=hospital+budget+&s2=&by1=n
23. Shohet Igal. Lavy. Sarel. Healthcare faciclits management: state of the art review vol 22 number 7/8 2004 210-22
24. Salonen. Antti, Deleryd. Mats, (2011),"Cost of poor maintenance: A concept for maintenance performance improvement", Journal of Quality in Maintenance Engineering, Vol. 17 Iss: 1 pp. 63 - 73
25. Sherwin, D. (2000), A review of overall models for maintenance management, Journal of Quality in Maintenance Engineering, Vol. 6 No. 3, pp. 138-64.
26. Scheinkopf. Lisa.J.1999. Thinking for change Putting the TOC Thinking Process To Use. APICS
27. Rosqvist, T., Laakso, K. and Reunanen, M. (2009), “Value-driven maintenance planning for a production plant”, Reliability Engineering & Safety System, Vol. 94, pp. 97-110.
28. Taylor III, Lloyd , Poyner, (2008),"Goldratt's thinking process applied to the problems associated with trained employee retention in a highly competitive labor market", Journal of European Industrial Training, Vol. 32 Issue: 7 pp. 594 – 608
29. Todinov, M.T. (2006), “Reliability analysis based on the losses from failures”, Reliability Analysis, Vol. 26 No. 2, pp. 311-35.
30. Vasseur, D. and Llory, M. (1999), “International survey on PSA figures of merit”, Reliability Engineering & System Safety, Vol. 66 No. 3, pp. 261-74.
31. Wireman, T. (1998), Developing Performance Indicators For Managing Maintenance, Industrial. Press, New York, NY

Fault Diagnostics for Railway Point Machines

Marius Vileiniskis, Rasa Remenyte-Prescott, Dovile Rama and John Andrews
Nottingham Transportation Engineering Centre, University of Nottingham, Nottingham, UK

Abstract

An increasing demand for railway transportation requires railway systems to be fault tolerant and safe. One of the main causes of train delays are point machine failures, which are usually associated with wear or misalignment of components. Such failures can be very hazardous and disruptive to the traffic on the network. In order to reduce failure effects there is an increasing emphasis in research on algorithms to detect, classify and prevent point machine failures. Commonly, condition monitoring systems on railway points are used, together with an alarm system based on threshold techniques. A more advanced condition monitoring system is needed which would include a fault detection and diagnosis system, so that the deteriorating condition could be detected prior to failure and failure consequences could be minimised.

In this paper a fault diagnostics methodology based on the method of one-class support vector machines (OCSVM) is presented. The aim of this study is to develop an online condition monitoring system for railway point machines, which would identify the deteriorating state of point machine prior to failure occurrence and potential causes of failures using measurements of current over time. The paper will report on the initial results of the study, when the OCSVM method is used to solve a pattern classification problem and distinguish between good and failed states of the system. The proposed methodology will be illustrated using field failure and maintenance data.

1. Introduction

According to the Office of Rail Regulation, around 1,450 million passenger journeys and 273,870 freight train movements were made from April 2011 till March 2012 [1]. With such high use of the railway infrastructure, any failure of a railway point machine might cause long delays, cancellations or even serious fatalities, such as the train derailment at the Potters Bar in 2002. For example, in the UK around 15% of train delays are caused by failures of switches and crossings (S&C), where the majority of S&C failures are caused by point machine failures [2]. To reduce the risk of point machine failure, a preventive remote condition monitoring (RCM) system should be in place. Such a system would give the information about the deteriorating condition of the point machine and detect failures in their earliest stage. Based on this information, engineers could plan the inspection and carry out the repairs at a more convenient time, for example, when the traffic volume is low.

An increase of the research interest in creating an online RCM system can be seen from a number of different methods proposed to detect point machine failures. The proposed models for fault diagnostics of railway point machines

from the trends of current can be divided into two major groups: models based on the laboratory testing bed data and models based on the data collected from the in-service point machines. One of the first approaches to fault detection was made by Oyebande and Renfrew, by using the statistical analysis methods [3], however, the distinction of failed state and working condition was not possible. Marquez et al. suggested analysing absolute values of the difference between the actual power data and reference power data [4]. Three criterions were employed to look at the curvature of the data: whether the shape was irregular, whether the time position of the maximum used force was within certain interval and whether the whole shape was symmetric with respect to the position of maximum used force. Authors showed that their proposed approach could detect 100% of the faults. However, this approach was tested using data from the laboratory point machine only and no performance results using real data were published. Several other methods have been proposed by Marquez et al.: spectral analysis [5], unobserved component models [6], time series analysis with auto-regression models [7] and principal component analysis [8, 9]. Very good results were achieved using some of these methods; however, their performance was not tested with the data from the in-service point machines.

When working with real data, Chamroukhi et al. proposed a new approach for fault detection of railway point machines [10]. The authors considered fault detection as a classification problem, i.e. to classify a given trend of current into one of the three classes – a trend when the movement is good, a trend with a minor fault or a major fault. Raw data was pre-processed to extract the main features of the trend shape. Then the authors used the mixture distribution modelling to learn the parameters of each class via the expectation-maximization algorithm [10]. Finally the movements were classified using a maximum a-posteriori rule, achieving a high rate of success (95% rate of classification). Further work on point machines by Chamroukhi incorporated regression models with a hidden logistic process to extract the features of raw data [11]. Authors also managed to employ the prediction of the point machine health state in their further research [12]. Results obtained using the classification algorithms with real data were very good, and their effectiveness could be further tested when the trends between the three classes are difficult to distinguish.

Based on good results discussed above, fault detection of point machines in this paper is also treated as a classification problem. In this paper, the one class support vector machine (OCSVM) is used for railway point machine fault detection, based on the measurements of current. The OCSVM, proposed by Scholkopf et al. [13], is an extension of the support vector algorithm (SVM) [14] for the case of unsupervised learning, when only one-class data is available for training the model. The OCSVM is combined together with an elastic metric called time warp edit distance (TWED), instead of the Euclidean distance, as the most commonly used method for comparing two time series. The TWED overcomes the weaknesses of the Euclidean distance, allowing some elastic matching between any two time series. This is an important feature for comparing two time series which are out of phase, i.e. when phase duration between two movements are different, as commonly observed from

the field data. The proposed approach is illustrated using the data available from the in-service point machines used on a part of the national railway network. The initial study presented in this paper forms a part of RCM process based on the current trends of point machines.

The structure of this paper is as follows. Section 2 explains the field data and it describes a commonly used RCM system and the most frequent faults of point machines. Section 3 explains the main features of the data used for the analysis in this study. The proposed approach with OCSVM and TWED is introduced in Section 4. Application of the proposed methodology using the field data is given in Section 5. The conclusions and further work are presented in the last section.

2. Railway point machines

Switches and crossings in the UK, as described in [15] and shown in Figure 1, are electrical and mechanical installations enabling railway trains to be guided from one track to another.

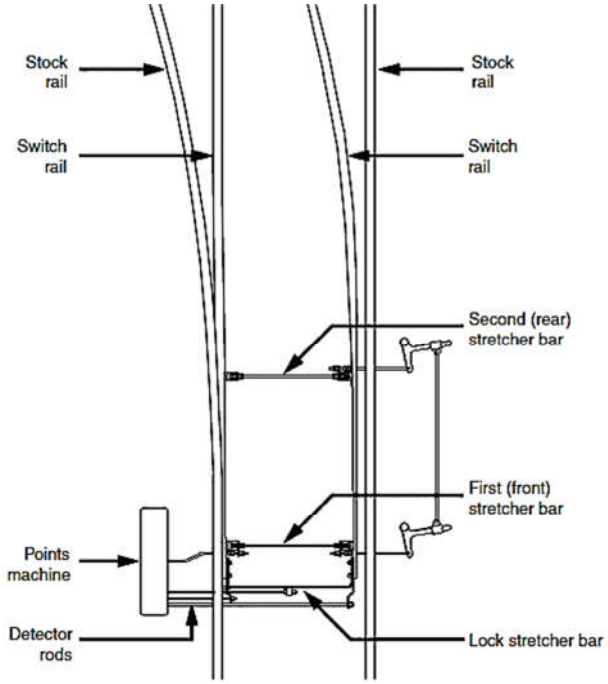


Figure 1. Common switch layout in the UK

The switch consists of a pair of linked movable switch rails (also known as points) lying between stock rails. A movement is carried out by a point machine. The points can be set in only two positions: normal and reverse. Two movement directions are identified: normal to reverse or reverse to normal. Location or state detection of the points is carried out by a two-position, polarised, magnetic stick contactor. The signal is fed back from these switches to a signal box where all point movements are controlled and monitored. Stretcher bars make sure that the switch rails remain at the correct distance apart. The number of stretcher bars can vary between installations depending on the curvature and the speed limit of the turnout.

The point machine considered in this paper is a clamp lock. It is a hydraulic point machine and its drive is provided by pumping oil through a hydraulic circuit. In Figure 2, electric current trends of operating phases of the clamp lock are shown. The first phase is the motor start up and inrush. The second phase is the unlocking of the switch blades. The longest operational phase is the switch blade movement. The last phase is locking of the switch blade. Due to a slight delay of the detection of locking the motor keeps running after the points have completed the movement. This is common to all clamp lock machines and thus a 'shark fin' at the end of the operation is observed, as shown in Figure 2.

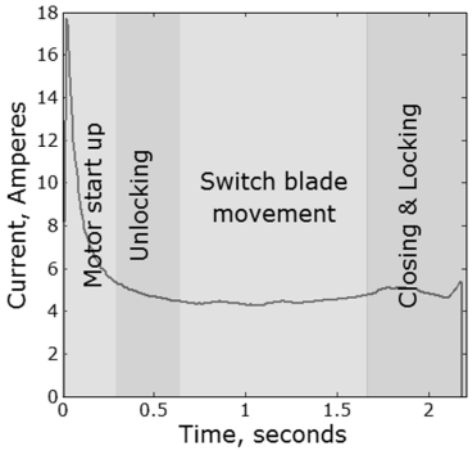


Figure 2. Operating phases of the clamp lock

2.1. A commonly used approach to the RCM system on point machines

When an online condition monitoring system for point machines is used, it consists of several parts: sensors, loggers, modems and servers. Current data is collected from the point machine via a non-intrusive system, i.e. current transducers are fitted to the point motor feed cable and they do not affect the operation of the equipment under observation. Loggers are used to collect the data from the point machine as current versus time measurement. It also logs the time and the direction of the movement. The data are further passed to a server that allows engineers to view the data and trends. Simple alarm thresholds for fault detection are implemented in the server. The values of the alarm levels must be tuned from time to time due to the seasonal changes and the wear out of the equipment. Commonly, such an approach based on threshold values is incapable to detect fault at its earliest stage.

The field data used in this study consists of measurements of current (in amperes) over time (in seconds), the date and time when the movement started and when it finished, the point machine ID and the direction of movement (normal to reverse or reverse to normal). Also, the information about faults logged, repairs or adjustments made, failure causes, failure and repair times is used in the study. Movements recorded between the date of

fault occurrence and fault rectification are labelled as faulty, otherwise they are labelled as good. The movement data collected one month before the failure event is considered in the analysis.

2.2. Failure modes of clamp locks

The common failure modes of clamp locks were described in [16]. The main failure modes identified were:

1. Detection is lost or out of adjustment;
2. Dry slide chair or poor lubrication;
3. Drive arm is out of adjustment;
4. Lock is out of adjustment.

Two failure modes are considered in this paper: dry slide chairs and dry lock arms and slide chairs. According to engineers, dry or contaminated slide chairs should show an increase of the current used throughout the phase of switch blade movement. Movements that were labelled as faulty and their failure causes were dry slide chairs and dry slide chairs and lock arms (in solid line, on the left and on the right respectively) are plotted against good movements (dotted line) in Figure 3.

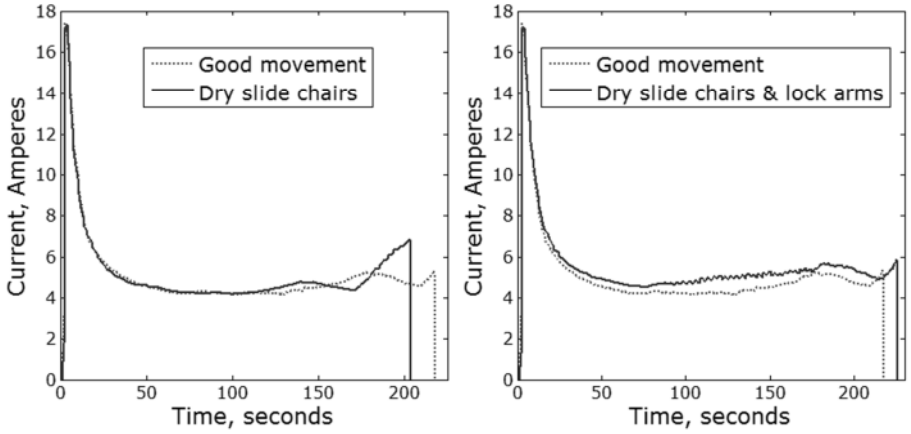


Figure 3. Dry slide chairs failure (left) and dry slide chairs & lock arms failure (right)

It can be seen on the left hand side of Figure 3 that there is an excess of energy used during the closing and locking phase. However, an excess of energy that should be used throughout the movement is not observed in this example. When the second failure is considered (dry slide chairs and lock arms), an excess of energy used should occur throughout the movement phase due to the dry slide chairs, as in the previous failure, and during the closing and locking phase due to dry lock arms, as observed on the right hand side of Figure 3. The solid line, which represents the movement when failure is present, is clearly above the dotted line, which represents the good movement. This kind of faulty behaviour can be easily explained from a

practical point of view. Since the slide chairs are lacking lubrication, the clamp lock needs to push the blades through the slides exposed to higher friction. This takes more effort, thus the excess of energy used is observed.

3. Features of real data

A number of features have been observed in the field data used for fault diagnostics of point machines. They have been of great importance in choosing a suitable method for the analysis.

One feature of real data is that each point machine seems to have its own specific current trend and thus point machines of the same type cannot be grouped for the analysis. This issue has also been raised by a former BRB Signalling Standards engineer Malcolm Tunley [17]. Differences between movements labelled as good and movements labelled as faulty of the same point machine are significantly smaller than the differences between two good movements of two identical type point machines (e.g. two clamp locks). That might be due to a different age of point machines, different configuration of S&C, different usage frequency etc. Thus data sets of individual point machines have to be analysed separately and there are far more trends of good movements than trends of faulty movements. Due to the huge imbalance of the data available in each class, the traditional two-class supervised learning classifiers are not suitable. The one-class classifiers should be used, which are trained only on the data of good movements to detect the abnormal movements. The OCSVM method is chosen in this study, since it has proven to give good classification rates in a wide area of fault diagnostics applications [18, 19] or data classification problems [20, 21].

The other feature is that the raw data for any two movements differs in length and thus the standard Euclidean distance cannot be used to compare the time series without modifying them. Some techniques can be used to transform the data in order to make it usable with the Euclidean distance:

1. Adding trailing zeros to the current time series which are shorter in length than the longest one available or some predefined length.
2. Dividing the current into a fixed number of segments and calculate the mean of the segments.
3. Rescaling the time series to make them of equal length [22].

Alternatively, there are several different similarity measures that are able to deal with the varying length of time series: TWED [23], edit distance with real penalty [24], dynamic time warping [25] etc. The idea of these algorithms is that they are trying to find corresponding elements in the given time series by allowing a scaling of the time axis. Such methods can be applied without data pre-processing. The drawback is that such methods are more time-consuming than the Euclidean distance and their parameter values need to be determined on individual case basis. The TWED is chosen in this study, and its effectiveness is illustrated by comparing the results with the ones obtained using the Euclidean distance.

4. Proposed approach for the fault diagnostics

Taking into account the features of data available, the proposed approach for the fault diagnostics of railway point machines can be described in the following stages:

1. Inputting and pre-processing the data.
2. Calculating the similarities between the movements using the TWED method.
3. Classifying the movements with the OCSVM to detect the abnormal movements.

More detailed explanation of each step of the proposed approach is given in the following subsections.

4.1. Data pre-processing

The first 300 ms of the movement measurements of the current is trimmed off due to the motor inrush. To filter out the noise and make the time series smoother, the time series are smoothed with an exponential smoothing:

$$A_j^S = A_{j-1}^S + \alpha\left(A_j - A_{j-1}^S\right), j = 2,3,\dots,n,\ A_1^S = A_1, 0 < \alpha < 1 \qquad (1)$$

where A_j is the *j*th data point of the original time series, A_j^S is the *j*th data point of the smoothed time series, α is the smoothing factor.

The smoothed time series are then used to calculate the similarity with the TWED.

4.2. Time warp edit distance

Marteau suggested a novel metric for time series matching [23], called time warp edit distance (TWED), which is similar to the edit distance with real penalties method [24], but it takes into account the difference of time when two data points are measured. Allocation of penalties for the time differences is based on a mechanical spring idea: the penalty is proportional to the length of the compared time interval. This proportional difference can be controlled with a stiffness parameter, ν_T. The TWED metric operates using three different operations: deleteA, deleteB and match. For the deleteA operation a penalty is added for deleting the data point in the first of two compared time series, for deleteB - for the data point in the second time series respectively. For the match operation a penalty is added based on the difference between two data points. Further details of the TWED algorithm can be obtained in [23].

The matching of two time series with TWED algorithm can be illustrated with a simple example. A copy of a good movement current trend was made by replicating its first element a number of times. The replicated current trend is then compared with the original trend. The alignment of these two time series given by TWED is shown in Figure 4. To visualize the alignment given by the

TWED, the replicated movement was shifted up in the plot. The leaning lines show the alignment by TWED. As can be seen from Figure 4, the TWED allows elastic matching of two time series and manages to deal with the shifted time series.

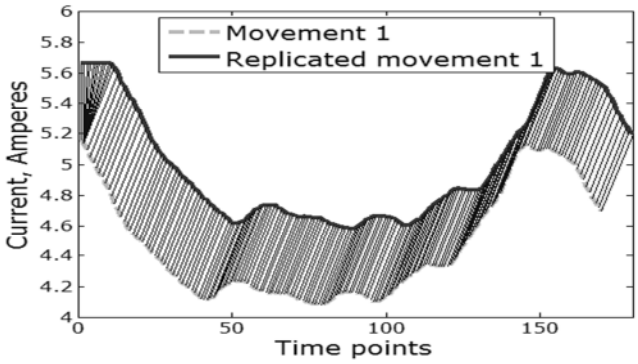


Figure 4. Alignment with TWED of two out of phase trends

An advantage of the TWED is that it is a metric and thus can be used with the support vector machines as a distance function for the Gaussian kernel instead of the Euclidean distance, as explained in the further section. However, TWED takes $O(n^2)$ computational time compared to the Euclidean distance $O(n)$. An idea from the Dynamic Time Warping method [26] can be applied to speed up the TWED calculations. Instead of comparing two very distant elements in the time series, we can apply Sakoe-Chiba band or Itakura parallelogram [26]. These two bounding methods restrict the comparison of the elements that are further apart than a pre-specified value. The banding of the TWED is also useful for practical purposes. The comparison of two data points in the time series which are far away from one another might even result in comparing the data points from two different operating phases of the point machine. The windowing of the TWED helps to avoid this.

4.3. One-class support vector machines

Scholkopf et. al. [13] suggested a new modification of SVM [14] to make it feasible to work with the unlabelled data. The idea is very similar to the binary SVM, but instead of trying to find the maximum margin between two classes, the OCSVM tries to separate the outliers from a given subset. Scholkopf et. al. developed an algorithm that creates a function f such that it returns the value of +1 for the most of the data points in a small region and the value of -1 for the other points, considered as outliers. Such outliers are separated from the origin with the maximum margin. The basic idea of the OCSVM is shown in Figure 5. The OCSVM is given only good class data for training as shown on the left hand side of Figure 5. The triangles represent support vectors which are used to calculate the separating boundary of the OCSVM, represented by a solid line surface. The support vectors are chosen from the training dataset in the training phase. In the testing phase of the OCSVM, all the points that fall out of the region defined by the separating boundary,

obtained in the training phase, are treated as abnormal data or outliers in the OCSVM, as shown on the right hand side of Figure 5.

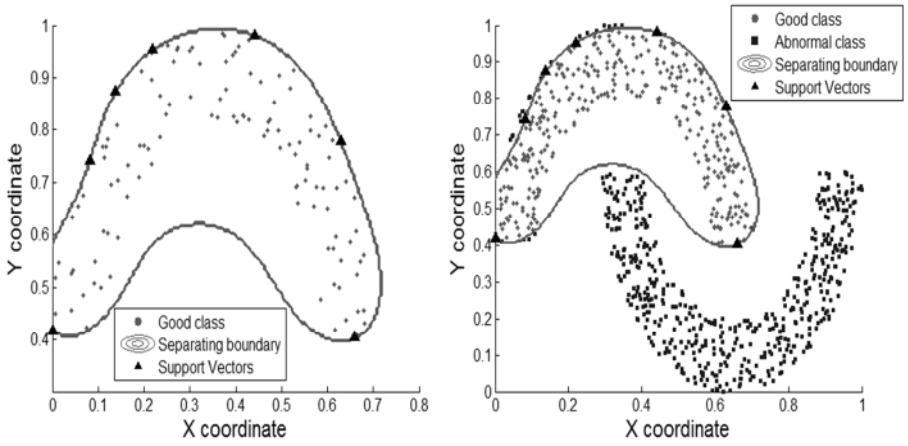


Figure 5. Training (left) and testing (right) phase of the OCSVM for horseshoe data

One of the main components of the OCSVM is a kernel function. The kernel function allows mapping the data into a feature space, where the separating boundary between normal and abnormal data is formed. OCSVM with a Gaussian kernel is used in this paper:

$$k(x, y) = e^{-\gamma D^2(x,y)}, \tag{2}$$

where $k(x, y)$ is the kernel function, γ is the width parameter, $D(x, y)$ is the selected distance function, x and y are the feature vectors.

The separating boundary of the OCSVM also depends on a parameter ν which is an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors. Classification rates of the OCSVM for a given data can be increased by changing the values of parameter ν and Gaussian kernel width parameter γ. The most common distance function $D(x, y)$ in the Gaussian kernel is Euclidean distance. However, due to the data features explained in the section 3, TWED is used instead of Euclidean distance. Further details of the OCSVM algorithm can be obtained in Scholkopf et. al. [13].

5. Application of the method using the field data

To illustrate the proposed approach in this paper, a simple example is considered, using chosen clamplock activity data of three months. 100 movements from normal to reverse direction were recorded during that period. 5 movements were labelled as faulty and 95 movements were labelled as good. Two faults were recorded and their causes were identified as dry slide chairs and dry slide chairs and lock arms respectively. The expected behaviour of these faults was described in section 2.2. Classification accuracy of the approach obtained with the TWED is compared with the results obtained using the Euclidean distance.

5.1. Training the model

The data for training and testing the model was divided in the following way: 20%, 30%, 40% or 50% of the total movements were used for training and the rest, i.e. 80%, 70%, 60% or 50% of the moments were used for testing. The data was pre-processed, as explained in Section 4.1. The TWED was used to compute the differences between the trends of each movement. Different combinations of TWED and OCSVM parameter values were chosen for the study Parameter values were chosen from the following ranges:

- TWED bounding window $r = (1,2,3,4,5,6,7,8,9,10,20,30,40,50,60)$.
- TWED stiffness parameter $\nu_T = 10^k, k = -5,-4,-3,-2$.
- TWED constant gap penalty $g = (0.05, 0.1,0.15,0.2,0.25,0.3)$.
- OCSVM parameter $\nu = (0.01,0.02,\dots,0.1)$.
- OCSVM kernel width parameter $\gamma = 10^k, k = -5,-4,\dots,0$.

The small values of parameter ν were chosen because point machine failures occur rarely, as mentioned in section 3. LIBSVM version 3.16 for MATLAB [27] is used as the core software for the OCSVM training and testing with some additional coding for the input and output of the data.

5.2. Testing the model

The efficiency of the OCSVM was tested with the movements, which have not been used in the training phase. A simple grid search (testing all the possible parameter combinations) was performed to find the optimal values for the TWED parameters r, ν_T and g, and OCSVM parameters ν and γ. The optimal ones were determined by following these rules:

1. The classification rate of good movements is higher than 80%.
2. The best classification rate of faulty movements is achieved.
3. The number of support vectors used is minimal.

Approximately 0.1 million of training and testing runs of the OCSVM have been made to test all the possible combinations of parameter values given above. In order to compare the performance of the TWED using different values of the parameters, all pre-calculated distances were normed to be in the [0;1] interval in the following way:

$$D' = \frac{D - D_{min}}{D_{max} - D_{min}}, \tag{3}$$

where D is the calculated difference, D_{min} is the minimal distance in all dataset, D_{max} is the maximum distance in all dataset, D' is the normed distance.

The best results with the optimal parameter values are summarized in Table 11.

Training size, %	γ	ν	r	g	ν_T	Accuracy, %	Good as good	Good as faulty	Faulty as good	Faulty as faulty
20	10^{-4}	0.06	4	0.1	10^{-2}	81.25	61	14	1	4
30	10^{-5}	0.04	5	0.1	10^{-2}	81.43	53	12	1	4
40	10^{-5}	0.03	6	0.1	10^{-2}	81.67	45	10	1	4
50	1	0.1	40	0.15	10^{-5}	80.00	37	8	2	3

Table 1. Best results achieved with different training data sizes

The classification rates are very similar for all the training data sizes. The best classification accuracy 81.67% (49 out of 60 tested movements were classified correctly) was achieved when 40% of the data was used for training the model and 60% for testing. The classification rates around 80% seem quite promising. However, a closer look at the results to determine the reasons why a higher success rate has not been achieved is needed. For example, the decision values of the OCSVM, when 20% of the data are used for training and 80% for the testing, are plotted in Figure 6. The circles represent the correctly classified currents, the x's represent the faulty currents classified as good and the triangles represent the good currents classified as faulty. Decision values that are above the solid line indicate that the movements were classified as good and below the solid line as faulty respectively. The decision values that are very close to the solid line indicate that the movements are very close to the separation boundary of the OCSVM. From Figure 6, one can see that quite a few good movements were close to the decision boundary (solid line) and some of the good movements were even classified as faulty. Further investigation is therefore needed to find out which movements were labelled as faulty in the testing data.

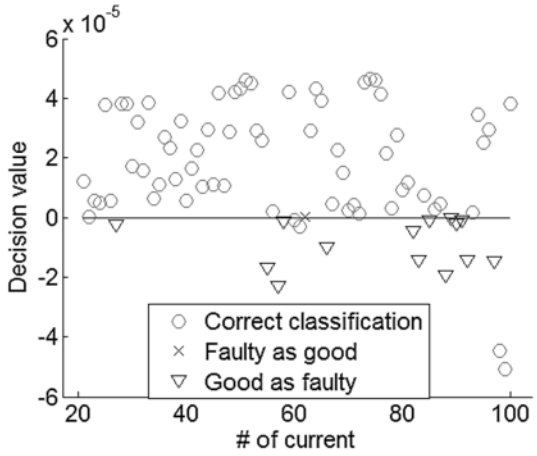


Figure 6. Decision values for the classified currents

From the five movements that were labelled as faulty, two movements are very similar to the good movements as can be seen in Figure 7.

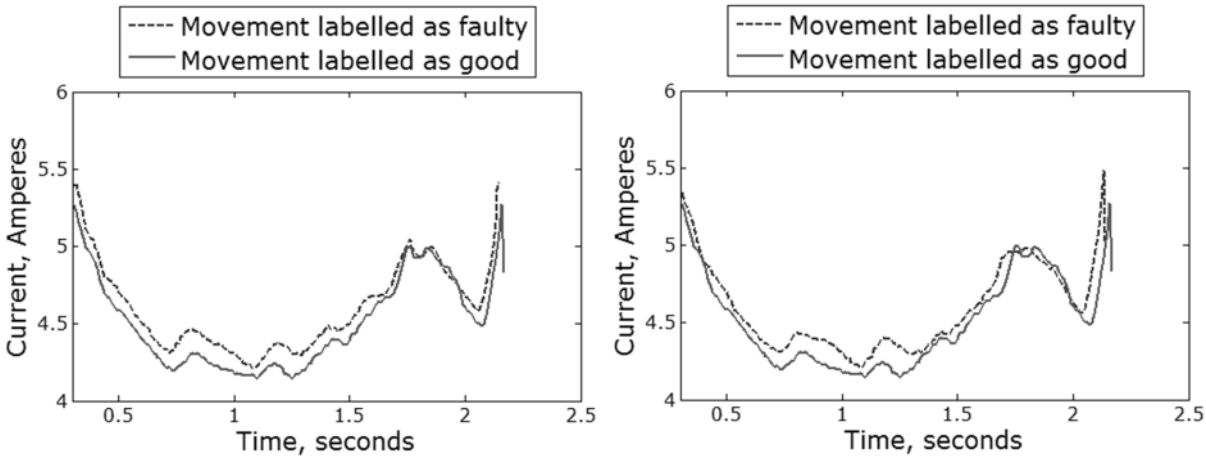


Figure 7. Movements labelled as faulty according to FMS database

The same can be said about the good movements. Good movements that look very similar to the faulty movements are given in Figure 8. The dotted line represents the faulty movements that were identified as dry slide chairs in the FMS database. The solid line represents good movements.

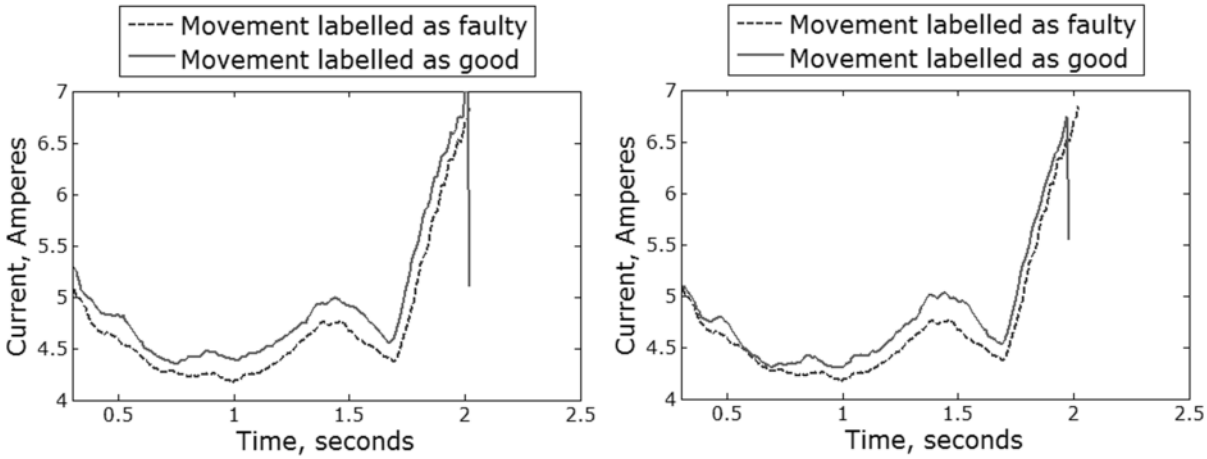


Figure 8. Movements labelled as good according to FMS database

These potential inconsistencies in automated labelling of the data can influence the behaviour of the OCSVM noticeably. The model was tested after re-labelling the movements given in Figure 7 and Figure 8, so that similar trends belong to the same class. The new decision values of the OCSVM are plotted in Figure 9. The classification accuracy has increased to 95% and all the faulty movements have been classified correctly. However, the decision values, starting from the 80th movement until the 98th movement, at which the actual fault occurred, are very close to the separation boundary and some of the good movements are classified as faulty.

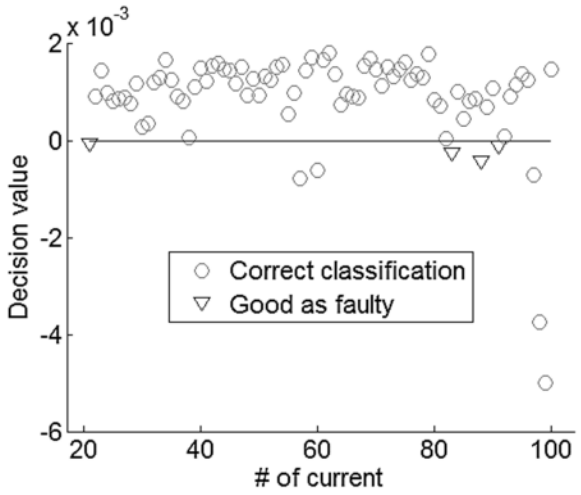


Figure 9. Decision values for the relabelled currents

The good movements that were classified as faulty and occurred prior to the fault recorded are plotted against some training data in Figure 10.

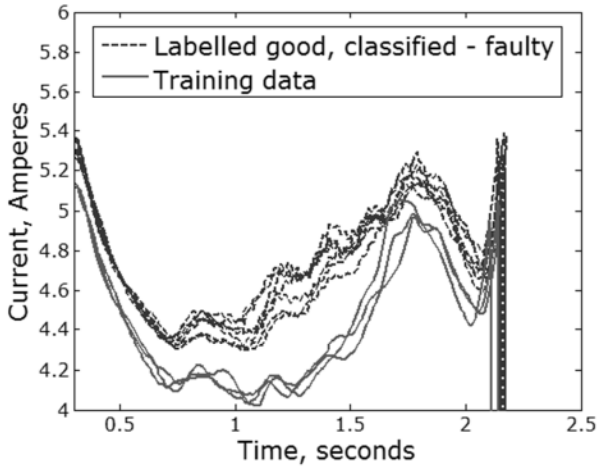


Figure 10. Comparison between the movements labelled as good

From Figure 10 one can notice that a big difference in energy used throughout the movement is observed in the movements that were labelled as good (dotted line) but subsequent movements led to the failure of point machine. The ability to identify such changes in the trends is the advantage of the method proposed in this paper. If further tests were successful, the method would be able to overcome the main disadvantage of the commonly used RCM systems – when a threshold value is used to raise an alarm and early signs of failure cannot be detected.

5.2.1. Comparison of TWED and Euclidean distance

In this section the results of the OCSVM with the TWED and the OCSVM with the Euclidean distance are compared. The raw time series of the measurements of current differ in length and the Euclidean distance cannot be used directly. For the analysis, the data were pre-processed using a uniform scaling method as suggested in [22]. All the trends were scaled up or

down to be of 200 points length (approximately an average number of data points, when the first 300 ms are trimmed off) according to the formula:

$$A'_j = A_{\lceil j \cdot \frac{n}{200} \rceil}, \qquad j = 1,2,\dots,n \tag{4}$$

where A_j is the *j*th data point of the original time series, A'_j is the *j*th data point of the scaled time series, n is the length of the time series A, $\lceil j \rceil$ is the ceiling function of j.

An illustration for two consequent movements of the point machine before and after uniform rescaling is shown on the left and right hand sides of Figure 11, respectively.

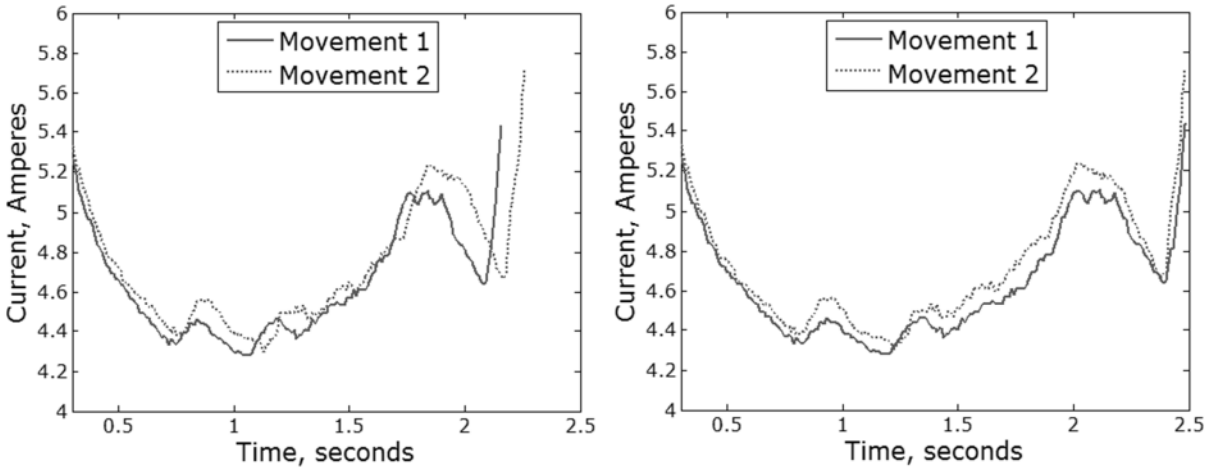


Figure 11. Unequal length (left) and uniformly scaled equal length (right) time series

The optimal parameter values for the OCSVM were again found with the grid search, as for the case with the TWED, described in section 5.2.

Distance	γ	ν	r	g	ν_T	Accuracy, %	Good as good	Good as faulty	Faulty as good	Faulty as faulty
TWED	0.01	0.09	2	0.3	10^{-5}	95	71	4	0	5
Euclidean	0.1	0.08	-	-	-	88.75	66	9	0	5

Table 2. Comparison of the results of OCSVM with TWED and OCSVM with Euclidean distance

The OCSVM with the Euclidean distance performed worse than the OCSVM with the TWED (88.75% compared to 95%). However, in terms of computational time, the Euclidean distance needs only $O(n)$ operations compared to $O(n^2)$ for TWED and this should be taken into account when online fault detection is considered.

5.3. Conclusions and future work

In this paper a fault detection method for railway point operating equipment based on one-class support vector machines is presented. When the field data is used, this novel method can distinguish between good and faulty movements of the point machine (95% classification rate). The time warp edit distance (TWED) is used to compare the trends of different lengths and its efficiency is compared to the traditional Euclidean distance (95% classification rate compared to 88.75%).

The main advantage of the proposed method is its ability to deal with the comparison of two time series when they are variously shifted along the time scale and differ in length, which is one of the main features observed in the field data. It also has the potential to identify the state of the system before the failure occurs, but its credibility is still to be tested on different failures and larger sets.

Further work is needed to address the discrepancies in the data, when faulty movements appear to be labelled as good and good movements – as faulty. This might overcome the issue of small samples for the analysis and increase the accuracy of detection, as preliminary tested in this paper. Further work could also include testing the speed of the approach and classifying faults by nature and their causes. If the method proves to be able to detect fault conditions prior to failure occurrence, it could be proposed as a part of the RCM system for railway point machines.

Acknowledgements

Rasa Remenyte-Prescott is The LRF Lecturer in Risk and Reliability Engineering. Dovile Rama is the Network Rail Research Fellow in Asset management. John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is also Director of The Lloyd's Register Foundation (LRF)[1] Centre for Risk and Reliability Engineering at the University of Nottingham. They gratefully acknowledge the support of these organizations.

[1] The Lloyd's Register Foundation (LRF) supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

References

1. Office of Rail Regulations, *The National Rail Trends Portal*. 2013.
2. Cornish, A., E. Kassa, and R. Smith, *Investigation of failure statistics for switches and crossings in the UK*, in *Railway Engineering 2011*. 2011: London.
3. Oyebande, B.O. and A.C. Renfrew, *Condition monitoring of railway electric point machines.* Iee Proceedings-Electric Power Applications, 2002. **149**(6): p. 465-473.
4. Márquez, F.P.G., F. Schmid, and J.C. Collado, *A reliability centered approach to remote condition monitoring. A railway points case study.* Reliability Engineering & System Safety, 2003. **80**(1): p. 33-40.
5. Márquez, F.P.G., W. Paul, and C. Roberts, *Failure analysis and diagnostics for railway trackside equipment.* Engineering Failure Analysis, 2007. **14**(8): p. 1411-1426.
6. Márquez, F.P.G., D.J.P. Tercero, and F. Schmid, *Unobserved component models applied to the assessment of wear in railway points: A case study.* European Journal of Operational Research, 2007. **176**(3): p. 1703-1712.
7. Márquez, F.P.G., D.J. Pedregal, and C. Roberts, *Time series methods applied to failure prediction and detection.* Reliability Engineering & System Safety, 2010. **95**(6): p. 698-703.
8. Márquez, F.P.G. and I.P. Garcia-Pardo, *Principal Component Analysis Applied to Filtered Signals for Maintenance Management.* Quality and Reliability Engineering International, 2010. **26**(6): p. 523-527.
9. Márquez, F.P.G. and J.M. Chacón Muñoz, *A pattern recognition and data analysis method for maintenance management.* International Journal of Systems Science, 2010: p. 1-15.
10. Chamroukhi, F., A. Same, P. Aknin, and M. Antoni. *Switch mechanism diagnosis using a pattern recognition approach*. in *4th IET International Conference on Railway Condition Monitoring*. 2008.
11. Chamroukhi, F., A. Samé, and P. Aknin. *A probabilistic approach for the classification of railway switch operating states*. in *Sixth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*. 2009. Dublin, UK.
12. Chamroukhi, F., A. Samé, G. Govaert, and P. Aknin, *A dynamic probabilistic modeling of railway switches operating states*, in *Proceedings of the 9th World Congress on Railway Research*. 2011: Lille-France.
13. Scholkopf, B., J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson, *Estimating the support of a high-dimensional distribution.* Neural Computation, 2001. **13**(7): p. 1443-1471.
14. Vapnik, V.N., *The nature of statistical learning theory*. 1995: Springer-Verlag New York, Inc. . 188.
15. McHutchon, M.A., W.J. Staszewski, and F. Schmid, *Signal processing for remote condition monitoring of railway points.* Strain, 2005. **41**(2): p. 71-85.
16. INNOTRACK, *List of key parameters for switch and crossing monitoring*. 2006.

17. Tunley, M., *Points to Failure*, in *IRSE news*. 2010, Institution of Railway Signal Engineers. p. 5-8.
18. Sarmiento, T., S.J. Hong, and G.S. May, *Fault detection in reactive ion etching systems using one-class support vector machines.* 2005 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop: Advancing Semiconductor Manufacturing Excellence, 2005: p. 140-143.
19. Martinez-Rego, D., O. Fontenla-Romero, and A. Alonso-Betanzos, *Power Wind Mill Fault Detection via one-class v-SVM Vibration Signal Analysis.* 2011 International Joint Conference on Neural Networks (Ijcnn), 2011: p. 511-518.
20. Song, X.M., G. Iordanescu, and A.M. Wyrwicz, *One-class machine learning for brain activation detection.* 2007 Ieee Conference on Computer Vision and Pattern Recognition, Vols 1-8, 2007: p. 2622-2627.
21. Zhang, N. *A novel image annotation based on one-class SVM*. in *Computer Science & Education (ICCSE), 2012 7th International Conference on*. 2012.
22. Fu, A.W.C., E. Keogh, L.Y.H. Lau, C.A. Ratanamahatana, and R.C.W. Wong, *Scaling and time warping in time series querying.* Vldb Journal, 2008. **17**(4): p. 899-921.
23. Marteau, P.F., *Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009. **31**(2): p. 306-318.
24. Chen, L. and R. Ng, *On the marriage of Lp-norms and edit distance*, in *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*. 2004, VLDB Endowment: Toronto, Canada. p. 792-803.
25. Sakoe, H. and S. Chiba, *Dynamic-Programming Algorithm Optimization for Spoken Word Recognition.* Ieee Transactions on Acoustics Speech and Signal Processing, 1978. **26**(1): p. 43-49.
26. Keogh, E. and C.A. Ratanamahatana, *Exact indexing of dynamic time warping.* Knowledge and Information Systems, 2005. **7**(3): p. 358-386.
27. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Trans. Intell. Syst. Technol., 2011. **2**(3): p. 1-27.

Probabilistic Analysis of Renewable Heat Technologies

Adam Thirkill[1], Paul Rowley
CREST, Holywell Park
School of Electronic, Electrical and Systems Engineering
Loughborough University, LE11 3TU

Abstract

There is a currently a lack of understanding regarding the real-life performance of domestic-scale renewable heat technologies. This limited knowledge can make it difficult to specify appropriate technologies for particular use-cases in terms of whether the technology will meet user requirements or the extent to which users will benefit from incentive schemes such as the UK's new Renewable Heat Incentive (RHI).

This paper describes a probabilistic evidence-based analysis of the performance of domestic solar thermal systems (STS). Using data gathered during the Energy Saving Trust's recent field trial study, 30 installations from across the UK were evaluated, and performance variations arising from both technical and non-technical system factors were analysed. The impact of factors such as system configuration, volumetric hot water consumption and auxiliary input were quantified. Principal component analysis (PCA) was applied to explore the correlation between system performance and system variables. A candidate Bayesian network (BN) has been developed which facilitates an effective means of both diagnosis and prognosis of system performance.

The results show a complex interaction of performance impact factors such as water consumption patterns and auxiliary heater cycling, resulting in mean system yields of 265.39kWh/m^2/yr with a standard deviation of 97.46kWh/m^2/yr. The analysis comprises a valuable platform for the development of advanced control approaches based on a Bayesian learning strategy with the intention of maximising system yields for specific use-cases.

1. Introduction

In the spring of 2014 the UK's Renewable Heat Incentive (RHI) will be open to domestic properties. The RHI tariff is payable based on deemed heat generation [1]. However there is a lack of understanding and evidence about the real life performance of renewable heating technologies meaning that the deemed heat generation could be misrepresentative of the actual heat provided by these systems. A lack of understanding introduces uncertainty regarding the performance of renewable heating systems thus leading to a risk in investment in terms of: financial return (from tariffs such as the RHI and savings on bills); and carbon reduction (the successful attainment of government carbon targets).

Uncertainty in the performance of STS due to a complex interaction between technical and non-technical factors [2]. Technical factors relate to the system

configuration such as the collector performance, tank size and collector area and are shown to have an effect on the solar input of the system and the system efficiency [3–5]. Non-technical factors are related to the way in which human users interact with the system. This includes: daily domestic hot water (DHW) demand [6], [7]; the DHW profile, or pattern of use [7–10]; the auxiliary timing [11], [12] and set point and load temperatures [13].

This study aims to evaluate the variation in the performance of STS and where the sources of this variation lie. Data from the recent Energy Saving Trust (EST) solar thermal field trial [12] has been used to identify this variation and possible influencing factors have been explored using principal component analysis (PCA). PCA has not been used before in the evaluation of STS performance. Principal component analysis can be used to simplify a complex multivariate problem by reducing many variables to a few uncorrelated variables (principal components) that retain the majority of the variation in the original set of variables [14], [15]. PCA can be used to explain variability in a set of data [16]. A novel approach to evaluating system performance in the light of uncertainty is presented in the form of a candidate Bayesian network. It shows the causal links between system variables and the probabilistic effects they have on the performance of the system.

2. Variation in solar thermal performance

Due to the complex interaction between highly variable factors, both technical and non-technical, the performance of the solar thermal system (in terms of the annual solar input, solar fraction, system efficiency and performance ratio) also experiences variation. The frequency distributions of the system parameters and performance metrics for the thirty systems are shown below. Variability in the system parameters is due to the inherent randomness of the incident irradiation and that of human behaviour, which affects the DHW demand and auxiliary input.

2.1 Variation in the performance metrics

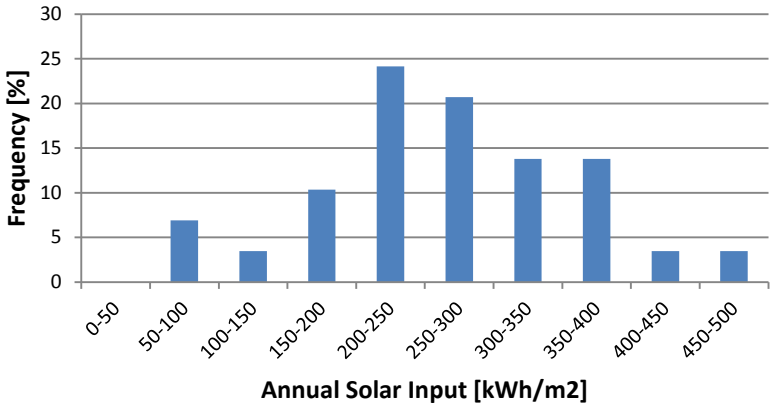


Figure 1. Distribution of the annual specific yield of the STS

Figure 1 shows the variability in the specific solar yield of the thirty STS. The majority of the systems show annual yields ranging from 50kWh/m^2 to

500kWh/m^2. The mean specific solar yield is 265.39kWh/m^2/yr with a standard deviation of 97.46kWh/m^2/yr.

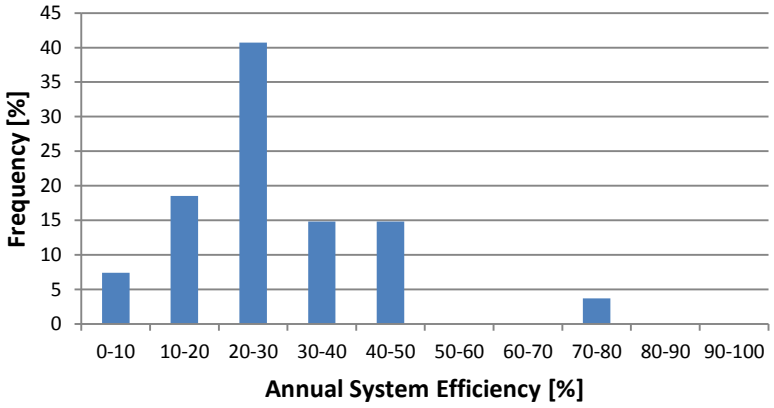


Figure 2. Distribution of the annual system efficiency

Figure 2 shows the distribution of the annual system efficiency. The mean efficiency of the systems is 27.5%. The system efficiency is given by (1):

$$\eta_{sys} = \frac{Q_{solar}}{I} \tag{1}$$

Where Q_{solar} is the solar input and I is the incident irradiation. For the annual efficiency this is calculated using annual solar input and annual incident irradiation. To calculate the average daily efficiency the daily efficiencies are calculated for the entire year and averaged over this time period.

The performance ratio is a measure of how close to the maximum efficiency obtainable at the given environmental conditions the system is operating. As with the performance ratio of photovoltaic systems it is a ratio of the actual yield to theoretical yield and can be calculated annually, monthly or daily [17]. Using the collector zero loss efficiency and heat loss coefficients the theoretical collector efficiency can be calculated using the irradiance and temperature difference between the collector and the ambient for each time step of measured data as in (2) [18]:

$$\eta_{theoretical} = \eta_0 - a_1\left(\frac{T_c - T_{amb}}{G}\right) - a_2\left(\frac{(T_c - T_{amb})^2}{G}\right) \tag{2}$$

This theoretical efficiency can be multiplied by the incident irradiation to give the theoretical solar input for that time step as in (3).

$$Q_{solar,theoretical} = \eta_{theoretical} I \tag{3}$$

This can be used in (4) to find the annual performance ratio.

$$PR = \frac{Q_{solar,actual}}{Q_{solar,theoretical}} \tag{4}$$

The performance ratio appears to have a high degree of variation from 10% to 90% with a mean value of 55.6%.

The solar fraction is defined as:

$$SF = \frac{Q_{solar}}{Q_{solar}+Q_{aux}} \quad (5)$$

Where Q_{aux} is the auxiliary input, which is the sum of the immersion and boiler inputs. Solar fraction ranges from 10% to 90% and has a mean value of 41.8%.

2.2 Variation in the key system parameters

The annual irradiation available to each system can range from 700kWh/m^2 to 1300kWh/m^2, with a mean of 1007kWh/m^2. This range is representative of irradiation levels in the UK [19]. The variation is due to different locations of the installations, orientations of the collectors and roof pitches.

The variability in DHW demand can be seen in Figure 3 and can be explained by the varying occupancies in the households as occupancy is shown to have the dominant effect on DHW demand [20]. The mean consumption is 111L/day, the Energy Saving Trust (EST) found mean consumption to be 122L/day +/-18L/day [20].

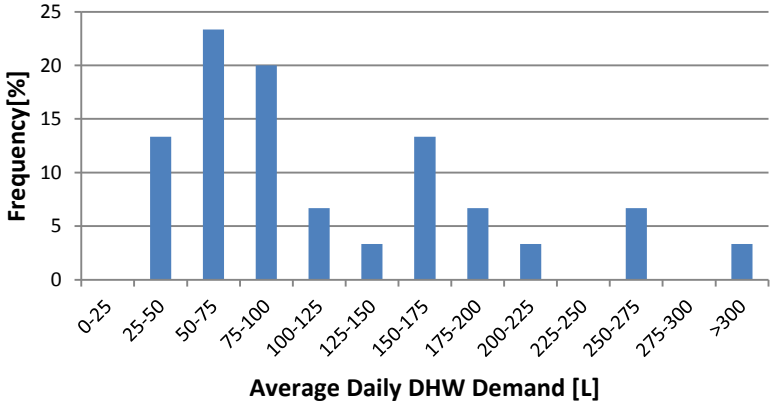


Figure 3. Distribution of average daily DHW demand

Figure 4 shows a variation in the annual auxiliary input. The average contribution from the boiler and immersion heater is 1949.7kWh/year.

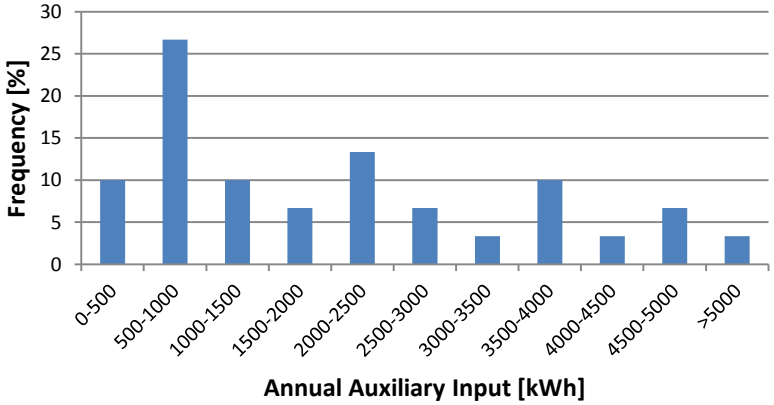


Figure 4. Distribution of the annual auxiliary input

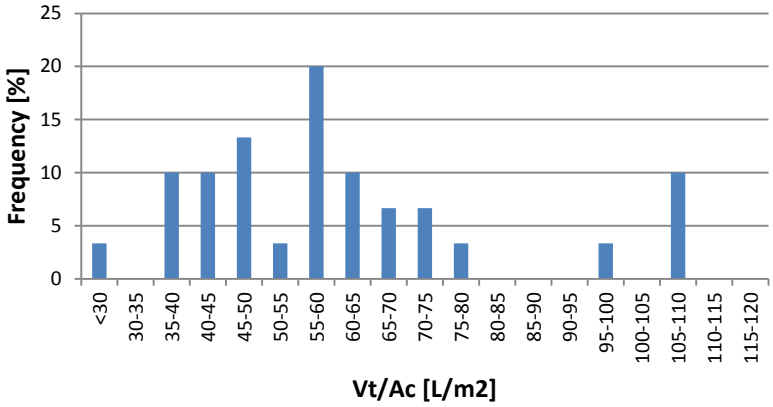


Figure 5. Distribution of the tank volume-to-collector area (VA) ratio

The variation in the VA ratio is shown in Figure 5. The average VA ratio is 62L/m^2. An optimum VA ratio has been shown to be between 50-75L/m^2 [4], [13].

Variation in the performance metrics is a result of the variation in key system parameters. PCA can be used to show which system parameters are responsible for the majority of variation in performance.

3. Principal component analysis (PCA)

The system parameters have complex interactions with each other therefore understanding which variables have the greatest effect on system performance can be difficult. PCA can be used to identify the most influencing variables allowing further analysis to be targeted. In PCA the original variables are transformed into new, uncorrelated variables called principal components (PCs) [14].

Figure 6 is a scree plot and shows the amount of variation in the data set represented by each PC. The first two PCs represent approximately 70% (67.3%) of the variation in the data. Between 80-90% of the variation should be accounted for in the selected number of PCs [14], therefore the first three PCs are used (82.2% of the variance).

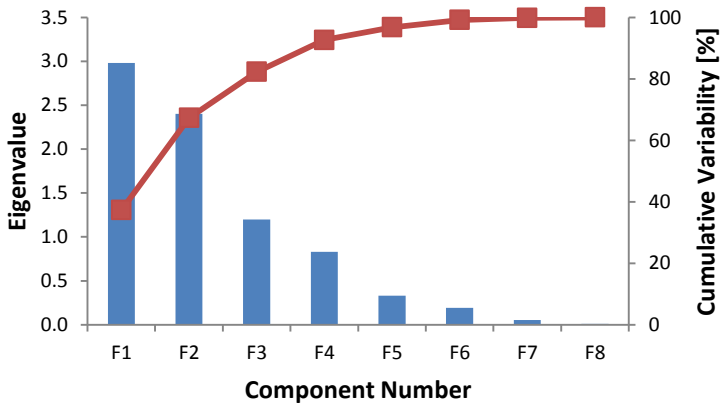


Figure 6. Scree plot demonstrating the variance

Variables	Q_{solar}	I	Q_{aux}	DHW	η_{sys}	SF	V/A	PR
Q_{solar}	**1**							
I	-0.038	**1**						
Q_{aux}	-0.027	-0.225	**1**					
DHW	0.292	-0.194	**0.803**	**1**				
η_{sys}	**0.945**	-0.264	0.051	0.412	**1**			
SF	0.320	0.072	**-0.777**	-0.394	0.334	**1**		
V/A	**0.612**	-0.097	0.172	0.177	**0.509**	-0.096	**1**	
PR	0.436	-0.285	-0.086	0.265	**0.590**	0.350	-0.018	**1**

Table 1. PCA Pearson correlation matrix

Table 1 shows the Pearson correlation matrix, used in PCA. It shows that there is significant correlation between the system efficiency and the solar input, which is to be expected since efficiency is a function of solar input. Interestingly there appears to be a weak negative correlation between irradiation and system efficiency and solar input. This suggests that there are other factors that have a greater effect on the solar input and efficiency of a system. The VA ratio seems to be one of these factors shown by a relatively strong positive correlation between this and solar input and system efficiency. It has been shown that as the VA ratio increases so does the collector efficiency and the solar input. The tank temperature decreases, which could explain the increase in solar input: A greater temperature difference between the tank and the solar coil allows for greater heat transfer between the solar heated water from the collector loop and the cooler water in the tank [13], [21]. There is a weak positive correlation between daily DHW demand and solar input. This suggests that although an increased DHW demand may increase the solar input due to re-charging the tank with cooler water [8], it may not be the governing factor in system performance. As mentioned previously it may be the pattern of use of the DHW that has a greater impact.

There is a strong positive correlation between the auxiliary input and the daily DHW demand. This is expected since more energy is required to heat larger draw offs to the desired temperature. There seems to be little correlation between the auxiliary input and the solar input suggesting that the two factors are not related. This is unexpected because when solar input is low, auxiliary input is expected to be higher to meet demand and vice versa. This relationship may not be visible on an annual time scale. Additionally, the timing of the auxiliary input may have a greater effect on solar input than the amount that is used. Poor control has been suggested to limit solar thermal performance [12], [11]. This is due to the auxiliary system competing with the solar collector. If the water in the tank has been heated by the auxiliary and has not been recharged with a sufficiently large water draw off before the solar resource is available, then the capacity for solar input is limited. The auxiliary input has the greatest effect on the solar fraction, greater than that of the solar input. This is represented by a strong negative correlation. This suggests that to improve the solar fraction the auxiliary input should be reduced. The way to reduce the auxiliary input seems to be by reducing the DHW demand.

Figure 7 shows the interaction of the variables based on the first three PCs (representing 82.2% of the variance). It clearly shows the strong negative correlation between solar fraction and auxiliary input as well as the close relationships between auxiliary input and DHW demand and solar input, system efficiency, performance ratio and VA ratio. The irradiation appears to have little effect on the other variables suggesting that the variation in the performance of the solar thermal system is governed more by the variation in the non-technical aspects of the system as well as the system configuration.

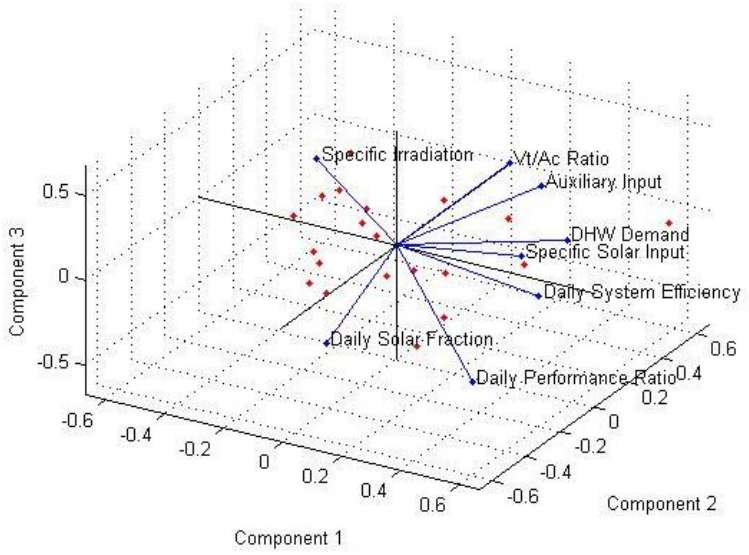


Figure 7. Correlations represented by the first three principal components

4. A candidate Bayesian network

The tariff levels in the RHI scheme are, in part, determined by the performance of renewable heat systems. Due to a lack of evidence about the performance of STS the appropriate tariff level is also an uncertainty. For STS the tariff is capped at 17p/kWh. It is capped because a higher tariff means that the same amount of renewable electrical energy could be produced from the cheapest source (wind). An evidence based probabilistic approach would be beneficial because the uncertainty in performance could be quantified. An appropriate tariff level may then be set with confidence or a capital grant may prove to be more suitable for STS installations.

The RHI tariffs are payable based on deemed heat generated by a system. The deemed heat will likely be based on the UK standard assessment procedure (SAP) [22], [23]. There are issues with using this approach that are largely concerned with the assumptions made and the exclusion of variability in certain parameters such as DHW demand and auxiliary input. By modelling performance probabilistically the STS may be targeted to households that are compatible with this technology. This evaluation could be performed by a Green Deal Assessor or Provider. By targeting STS at compatible households further uncertainty in tariff level setting might be reduced. A Bayesian network (BN) facilitates this.

The Bayesian network is a graphical tool that allows causal links between variables to be represented visually. BNs allow the incorporation of uncertainty into inferences made about the performance of STS. Prior knowledge from field trial, laboratory or modelled systems can be used to make estimates on likely performance [24]. BNs are resilient to limited data sets, which is ideal for STS performance estimations where data is often limited. BNs allow inferences to be made from parent to child and child to parent by virtue of Bayes Theorem [25]. The prognosis (parent to child) function is useful for designers of STS or for the implementation of tariff schemes such as the RHI. The diagnosis (child to parent) function is useful in targeting the likely variables causing sub-optimal performance.

Advanced control strategies that use Bayesian inference to learn user behaviour could be developed with the aim to provide effective control thus improving system efficiencies and solar thermal yields. The ability of BN to use probabilistic data for environmental effects and to learn and update probabilities for energy requirements of specific users means that incorporating this ability into a control system for the auxiliary heating would improve the behaviour of the whole system. System inputs such as solar and auxiliary input in the network may act as input signals to the control system. The energy required by the user is the desired output of the system. Future states of the solar input signal can be estimated through the use of evidence about this variable and may allow effective timing of the auxiliary to be performed in order to maximise solar contribution. Examples of controlling systems using Bayesian methods can be found in [26], [27].

4.1 Constructing the Bayesian network

There are two elements to a Bayesian network: the structure and the conditional probabilities of the variables. There are two ways in which the structure can be developed: Learning from raw data, which requires a rich source of data to be effective; or from literature and elicitation from experts [28]. Many software packages on the market allow structure learning such as HUGIN, Netica and BayesiaLab. However in the case of STS analysis the data has been found to be incomplete and highly variable and so structure learning is not practical. Therefore elicitation of the structure from the raw data is performed and literature is used to support this process.

Shipworth suggests realist synthesis as one way of developing the structure of a network. In this method questions or theories are developed and evidence, from literature or data, is found to support or reject the theory. The factors identified in the realist synthesis are evaluated in terms of the context of the study in which they were found and the direction of the effect between factors is determined. The strength of the evidence that supports the relationships between factors is considered and gives an indication of the confidence level. The variables included are discretised to a level that still represents the correlations between factors but the number of states of each variable is minimised to reduce the number of probabilities in the network. These steps develop the structure and initial conditional probabilities. There are similarities between this method and the method of constructing Bayesian

networks from causal maps [29]. This method focuses on forming the causal map (CM) from data and textual analysis of literature to determine the system variables and the direction of causality between them. This is similar to the realist synthesis but the output is a causal map, which is later transformed into a BN. The CM provides the structure of the network. The CM is modified by removing abductively reasoned cause/effect relationships (where the direction of cause and effect is misrepresented) and feedback loops (which are not allowed in BNs. The number of states of each variable are then determined. Conditional probabilities are then elicited from data or expert opinion where data is incomplete.

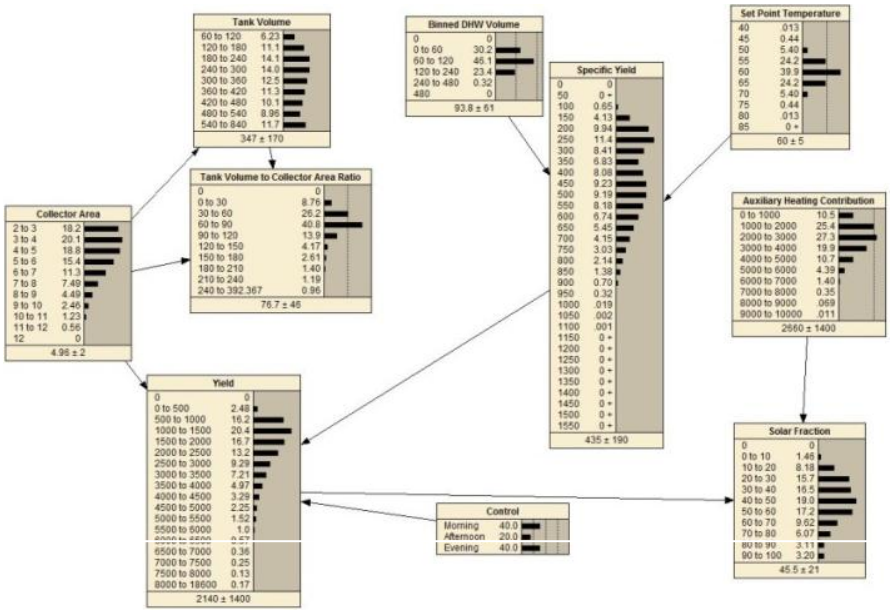


Figure 8. Candidate Bayesian network for STS

The candidate Bayesian network shown in Figure 8 was developed using a combination of the two methods outlined above. An initial review of literature identified the system variables to be included and a preliminary causal map was drawn from this. The data used to populate the network came from an EST DHW field trial [20] and the Viridian solar thermal field trial [11] as well as literature about STS performance. The development of the network is an iterative process, which allows for updated information to be included from new theories or data sets. Therefore the next stage is to revisit the network with the inclusion of the EST field trial data and refine the network structure and conditional probabilities in light of this new information.

4.2 Initial results using the candidate Bayesian network

Although this Bayesian network is in its initial stages it has been used to predict the annual solar input for an additional case study STS, results are shown in Figure 9.

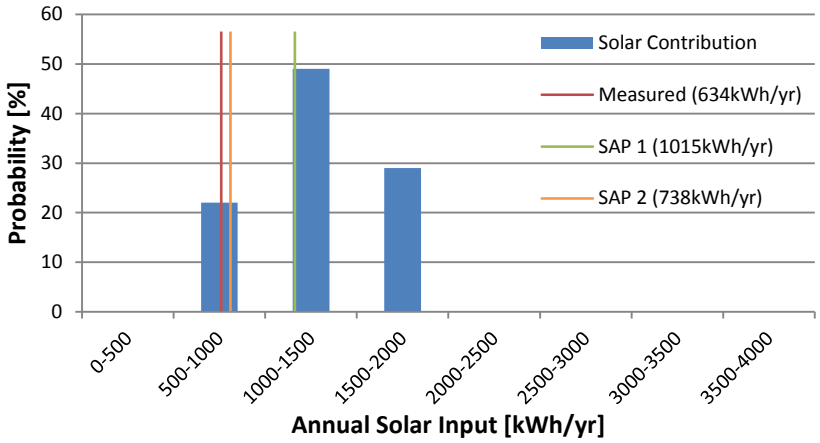


Figure 9. Predicted and measured annual solar input

SAP 1 refers to the calculations using the same assumptions SAP would make about annual irradiation and DHW demand based on the floor area of the house, whereas SAP 2 substitutes measured data in places where a SAP assumption would be made.

The results from the BN show a distribution of possible annual yields that a system of this type may achieve. In comparison, SAP predicts a single value for annual yield, which suggests that identical systems installed in similar households would achieve the same yield and that this yield could be expected every year. From analysis of field trial data this is not the case. Substituting measured data in the SAP calculation provides a closer approximation to the actual value of solar input. However this annual solar input can vary from year to year, which is not accounted for in the SAP estimation.

5. Conclusion

Following a PCA on the aforementioned variables it is seen that the majority of the variation in solar input is due to the variation in VA ratio. The variation in solar fraction is due to the variation in auxiliary input, which in turn is linked strongly to the DHW demand. The amount of DHW consumed and auxiliary input seem to have little effect on the solar input when evaluated on an annual basis, but the time of use in relation to when the solar resource is available and when the auxiliary is fired could have an impact which may lead to an improvement to the solar fraction if timed effectively. An investigation into this will form part of the future work. The system efficiency can be improved by careful consideration of the VA ratio which could increase the solar input.

The candidate Bayesian network presents a map of the relationships between variables that affect STS performance. It offers a practical means to manage and quantify uncertainty in STS performance. It provides a distribution of the probable performance of the system based on evidence contained in the nodes. This uncertainty is important to consider in systems such as STS since the inputs and outputs of the system vary greatly. For government to successfully implement the RHI for STS this variability and uncertainty must be accounted for to allow an appropriate tariff level to be established.

Uncertainty in the performance of these systems leads to an uncertainty in the appropriate tariff level and therefore in the return on investment for the user and whether they will be fully compensated for the initial cost of the system. This can determine the success of STS as part of the RHI scheme.

As future work the relationships identified in the PCA as well as the strength of them, as quantified by the Pearson correlation matrix, will be used in the refinement of the candidate Bayesian network structure. Further analysis of auxiliary timing will also be investigated using case studies selected from the EST solar thermal field trial. New evidence elicited from this analysis will be used to improve the confidence in the conditional probabilities of the variables.

References

[[1] EST, “Renewable Heat Incentive (RHI),” 2013. [Online]. Available: http://www.energysavingtrust.org.uk/Generating-energy/Getting-money-back/Renewable-Heat-Incentive-RHI. [Accessed: 24-Oct-2012].

[2] F. Domínguez-Muñoz, J. M. Cejudo-López, A. Carrillo-Andrés, and C. R. Ruivo, *Energy and Buildings*, **47** 474-484 (2012).

[3] M. Lundh, K. Zass, C. Wilhelms, K. Vajen, and U. Jordan, *Solar Energy*, **84** [7] 1095-1102 (2010).

[4] A. M. Shariah and G. O. G. Lof, *Renewable Energy*, **7 [**3] 289-300 (1996).

[5] M. C. Rodríguez-Hidalgo, P. A. Rodríguez-Aumente, A. Lecuona, M. Legrand, and R. Ventas, *Applied Energy* (2012).

[6] S. R. Allen, G. P. Hammond, H. A. Harajli, M. C. McManus, and A. B. Winnett, *Energy*, **35** [3] 1351-1362 (2010).

[7] S. Knudsen, *Solar Energy* **73** [1] 33-42 (2002).

[8] S. Furbo, E. Andersen, S. Knudsen, N. K. Vejen, and L. J. Shah, *Solar Energy*, **78** [2] 269-279 (2005).

[9] [9] U. Jordan and K. Vajen, *Solar Energy*, **69** 197-208 (2001).

[10] A. M. Shariah and G. O. G. Löf, *Solar Energy*, **60** [2] 119-126 (1997).

[11] D. Forward and C. Roberts, “Viridian Solar - Clearline Solar Thermal Field Trial,” 2008.

[12] EST, “Here comes the sun : a field trial of solar water heating systems,” 2011.

[13] A. M. Shariah and A. Ecevit, *Energy Conversion and Management*, **36** [5] 289-296 (1995).

[14] I. T. Jolliffe, *Principal Component Analysis*, 1st ed., Springer-Verlag New York Inc., (1986).

[15] MathWorks, “Feature Transformation,” 2013. [Online]. Available: http://www.mathworks.co.uk/help/stats/feature-transformation.html#f75476. [Accessed: 15-Oct-2012].

[16] J. E. Jackson, *A User’s Guide to Principal Components*, 1st ed., John Wiley & Sons Inc., (1991).

[17] J. Mondol, Y. Yohanis, M. Smyth, and B. Norton, *Energy Conversion and Management*, **47** [18–19] 2925-2947 (2006).

[18] German Solar Energy Society, *Planning and Installing Solar Thermal Systems*, 2nd ed., Earthscan, (2010).

[19] M. Šúri, T. a. Huld, E. D. Dunlop, and H. a. Ossenbrink, *Solar Energy*, **81** 10 1295-1305 (2007).

[20] EST, “Measurement of Domestic Hot Water Consumption in Dwellings,” (2008).

[21] K. Çomaklı, U. Çakır, M. Kaya, and K. Bakirci, *Energy Conversion and Management*, **63** 112-117 (2012).

[22] M. Crowther, H. Charlick, B. Bates, and R. Green, “Report to DECC on Heat metering for the RHI,” (2010).

[23] S. Kelly, D. Crawford-Brown, and M. G. Pollitt, *Renewable and Sustainable Energy Review*, **16** [9] 6861-6878 (2012).

[24] S. Washington and J. Oh, *Accident Analysis and Prevention*, **38** [2] 234-247 (2006).

[25] M. Lampis, “Application of Bayesian Belief Networks to System Fault Diagnostics,” Loughborough University, (2010).

[26] R. Deventer, J. Denzler, and H. Niemann, *Proceedings of the IBERAMIA/SBIA 2000 Workshops* (2000).

[27] M. Moulin, *32nd Annual Conference on IEEE Industrial Electronics* (2006).

[28] D. Shipworth, *ECEEE 2005 Summer Study,* 1381-1391 (2005).

[29] S. Nadkarni and P. P. Shenoy, *Decision Support Systems*, **38** [2] 259-281 (2004).

Quantifying Technical Risks: Insights into Theory-Practice Tensions in the Elicitation Process and Method

Gillian Anderson, Matthew Revie, Lesley Walls
Department of Management Science, University of Strathclyde, Glasgow, Scotland

Abstract

The elicitation of probabilities about possible future events to support technical risk assessment is challenging even when modelling specific projects. Yet probability elicitation is core to many organisation-wide technical risk management processes and will be conducted at regular intervals, across diverse sets of assets and by multiple engineers, perhaps across many heterogeneous sites. We investigate how such elicitation is conducted in practice to examine how the reality of what happens relates to the theoretical principles of subjective probability judgement elicitation. This paper reports a study conducted with a large organisation to help inform their technical risk management process. We report two inter-related strands of work. First an analysis of interviews with selected engineers who participate as experts to provide to probability judgements in the case organisation. Second, analysis of data collected through an experimental study designed to examine the robustness of several probability elicitation methods for samples of engineers from the case organisation and a control group of post-experience part-time students, all in full-time employment. Our interviews reveal multiple examples of the use of heuristics and evidence of bias that explain some disparities evident to the organisation. Our experimental data indicates that the choice of numerical and verbal descriptors on the scale has a major impact on the probability values elicited. The implications of our findings for the technical risk management process are discussed.

1. Introduction

This project was motivated by the opportunity to work with a large utility company which was reviewing its Technical Risk Management (TRM) process. TRM is the overarching framework the organisation uses to manage technical risk across multiple sites and serving several engineering disciplines. As well as assessing whether the current TRM process complied with the UK HSE's ALARP guidelines (HSE, 2001), the company was also reviewing the alignment of its TRM with that of its international parent organisation. Consequently there were opportunities to inform the further development of the company's already mature Enterprise Risk Management system (ERM) which supports the TRM process.

The elicitation of subjective risks (i.e. probabilities and consequences of diverse operational events) by multiple engineering experts is core to quantification in the TRM. The practical challenges of gathering judgements were already appreciated by the organisation and concerns had been expressed about the potential bias in estimated risk quantities in relation to perceived, credible risk priorities. A proposed change to assess risk by type of asset rather than at a plant level had the potential to make elicitation of risk even more challenging. This is because such a change implies a new way of framing elicitation of engineering judgement since multiple assets, although spatially separated, will be considered in the same risk assessment.

Although a large literature on the elicitation of structured expert judgement to support technical risk analysis exists (e.g. see O'Hagan et.al. 2006), there is little research into probability elicitation for a recurrent process of organisation-wide technical risk management. Hence it is interesting to investigate how such elicitation is conducted in practice so that we can examine the reality of what happens in relation to the theoretical principles of subjective probability judgement elicitation. Doing so should permit us to provide more meaningful recommendations about the further improvement of the elicitation process and method to technical risk management.

To meet this overarching goal, our study had two primary objectives:

- To investigate how typical engineering experts within the organisation actually make probability judgements as part of a technical risk assessment;
- To examine the performance of alternative methods for the elicitation of probabilities to better understand how they might affect judgements captured as part of a technical risk assessment.

In Section 2 we describe how we approached the interviews with engineering experts in the case organisation to find out how they made probability assessments and we discuss our main findings. Section 3 then describes the design and analysis of a matched experimental study to compare alternative methods for probability elicitation. This study includes the method currently used by the organisation and aims to compare it with other commonly advocated methods. We reflect upon the implications of our findings about both the process and method for elicitation within the case organisation, and more generally, in Section 4.

2. **Probability Elicitation through the Lens of the Case Organisation**

The standard procedure for calculating technical risk is to compute the product of probability and consequence, where each are measured on a 5 point scale hence providing a risk score bounded between 1 and 25. The probability scale is given between 1 (remote) to 5 (certain), where each verbal description translates to specific numerical probability ranges and is defined in company documentation. The consequence valuations are also based on a quantitative scale from 1 (minor injury) through to 5 (loss of life). In this article we focus upon the probability scales only, although consequence was also reviewed as part of the wider study.

2.1 Design of Case Interviews

Semi-structured interviews with nominated experts were the means used to investigate how such probabilities tend to be elicited in practice. The experts selected were identified by the a Senior Risk Analyst and were all engineers who were highly experienced and influential within the organisation as well as being recognised by their peers as being qualified to provide values for the risk score. The selected experts were from different engineering disciplines, different sites and all were managers with over 20 years of experience. Semi-structured interviews led by an independent researcher (the first author) were deemed to be an appropriate method to ask questions and explore issues around the following points:

1. To gain additional detail on current practice, especially issues of concern and good practice;
2. To establish how experts in the case organisation currently think about the role of probability elicitation in making risk assessments;
3. To identify any differences between experts in the manner in which they apply the standard method for eliciting probabilities;
4. To establish if any across site differences exist and, if so, to appreciate the nature of such differences;
5. To understand the type of information engineering experts draw upon to determine the probability of an event and how they think through the formation and expression of their probability assessments.

The questionnaire was designed to cover the above issues and was used in an informal way during the interviews as the researcher managed the discussion real-time. The interviews comprised three parts, background information on the engineers and their experience, general questions about how the TRM process worked in practice and, finally, a post-it session where the engineers provided an example risk assessment and demonstrated how they typically elicited a value for the probability of occurrence. Interviews took place over a period of 4 weeks in April 2012, three were face-to-face on-site and one by an electronic medium. Being on-site also allowed the researcher to obtain a tour of facilities and gain better understanding of the assets and environment in which probability assessments were being made. The interviews were recorded, transcribed and the information from them was used to prepare the findings.

2.2 Key Findings on Organisational-Wide Probability Elicitation in Practice

Many issues arose that provided new insights to the organisation.

While the importance of assessing probabilities was acknowledged, concerns were raised about the defensibility of the some judgements due to the *operational process*. A variety of reasons were articulated for these concerns. They include that experts had to consult separate documentation to get the quantitative and qualitative

descriptions of probability. In some documentation, different probability values were given for the same qualitative description, thus assessments might be inadvertently made on different scales. Experts recognised that the probability values provided fed into a larger analysis that impacts investment decisions of which their assets or site might be a beneficiary. Hence the opportunity to manipulate the elicitation process exists, even if there is no evidence that this was acted upon. Some probabilities will be made by a single expert, but others by a group within an open meeting with a lack of standard process for arriving at an aggregated value. The current process was regarded as complex and time-consuming with the transparency and traceability of probability values being lost during the chain of analysis. For example, multiple reviews of the top risks were conducted within the organisation allowing probability values provided by nominated experts to be questioned and, in some cases, revised without feedback to the original experts.

Experts generally found it challenging to articulate *how* to verbalise their thought process to arrive at a probability. It is apparent that some experts prefer to think in terms of number of occurrences of an event over a time period, such as a specified number of years, whereas other experts think about probability as a percentage of times an event might occur. In part, this might be natural since some assessments will be required for time based operational assets and others for probability of on-demand failures. The current scoring scale with 5 levels was considered too narrow, especially in regard to civil assets. The timing of *when* experts are asked to make a probability assessment was also regarded as important. For example, there was awareness that if it was after an incident then there would be a tendency to anchor on recent observational data more than usual. For each event there were often multiple aspects to consider and experts tended to structure these in different ways. For example, given the need to make an assessment for three assets that had operated for 20 years without incident, the one expert reasoned that this equated to 60 years without incident, while another believed that it represented 20 years without incident.

Issues were also raised in relation to *who* provided the probability assessments. Often there is only a single expert who was deemed qualified to make the assessment, making it difficult to validate or clarify the specification. While equipment operators might raise concerns about assets and provide feedback to engineering experts, the recognition of such concerns and rolling in this understanding into any probability assessment appears to depend on relationships, trust and intuition of the expert with the operator. More generally, it emerged, during interviews, that experts viewed probability quantification as not only an output of their own personal knowledge, but also dependent on the quality of their so-called knowledge network, both formal/informal and internal/external to the company. One concern expressed by the experts was that the nature of these knowledge networks was changing due to redundancies, an ageing workforce and a reduction in the ability to foster and develop these trust based networks. This was regarded as have

the potential to affect future availability of experts, expertise and the nature of uncertainties in making probability assessments.

2.3 Interpreting Findings in Relation to Scientific Principles of Elicitation

Cooke (1991) developed five principles of structured expert judgement. Namely: reproducibility (all calculations must be reproducible); accountability (source of expert subjective probabilities identified); empirical control (Expert assessments should be susceptible in principle to empirical control); neutrality (method encourages experts to state true opinions); and fairness (all experts are treated equally a-priori). It could be argued that the current company process breaks many of these principles. For example, calculations may not be reproducible as there is ambiguity between the qualitative definition and the probability values. As multiple experts can change the probability without a clear record being maintained, there is a lack of full accountability with the current system. Experts are aware the system is used to fund projects and so are not incentivised to give their true opinion. As there might be multiple changes to an original probability provided by an experts the true source might not be identifiable and implicitly all experts are not being fairly.

In their seminal book, Kahneman, Slovic and Tversky (1982) classify and discuss different heuristics and biases that might be expected in the elicitation of expert judgement. Many examples were evident in the narrative from the interviews. For example, educational bias is presented due to heterogeneity across the organisation in relation to the extent and use of knowledge networks. Anchoring was prevalent when experts explained that they ranked risks prior to assessing their probabilities and many assessments were made post-occurrence of an event; hence the availability of observational data gives rise to a bias on specification. Motivational bias might arise since multiple changes could be made to probability values after an expert had specified his/her belief. Since probabilities and consequences were to be input to the ERM system via the same input screen, there is the possibility of structural bias occurring. Even though a standard data capture process existed, the method to be used for probability elicitation by different experts lacked consistency of application. Since the method for probability elicitation is core to the entire technical risk process, we have examined it in more depth.

3. Experimental Study of Alternative Methods for Probability Elicitation

Through a designed experimental study we explore how the current method of expert judgement elicitation used in the case organisation compares with acknowledged alternatives. In particular, we investigate the efficacy of methods in terms of minimising opportunities for bias and perceptions of experts on ease-of-use.

To access our target population, the engineering experts at the case organisation, a sample frame was constructed with the assistance of the Senior Risk Analyst, and 180 representative engineers and managers were selected to take part (i.e. we now refer to these as the company group). To allow us to compare our findings with a

control group of ‘matched’ professionals, we invited 150 part-time, post-experience international MBA students at Strathclyde Business School to take part. All students were in full-time employment, many as engineers or in other technical disciplines, with the remainder in other business functions (i.e. we refer to these as the student group). Although not a perfect match to the company experts, as post-experience professionals the students represent a reasonable and accessible control group.

The study was conducted over the period July and August 2012. More details about its design follow.

3.1 Selection of Probability Methods

The methods selected for our study were informed by the case interview feedback and from the literature (Van Der Gaag, 1999). While it is possible to consider a larger set of methods, we constrained our study to five methods due largely to the practicalities of experimenting with real professionals. However the five methods selected included a mix of scale representation. Three methods used a visual scale with the other two methods required numerical values to be specified for either a direct statement of probability or about the uncertainty in terms of lower or upper bounds on the judged probability. Of the three methods presented visually, two had joint verbal and numerical descriptors, while the third had numerical values only. Verbal numerical (VN) scales are believed to have the advantage that they allow the expert to use a scale familiar to them and choose whether to use the words or numbers, which can be a fast way to elicit probabilities (Renooji, 2001). In summary, the five methods and their rationale are as follows:

1. Verbal numerical with 5 points (VN5pt) – Similar to the scale used in the organisation, differing only in that it had a single verbal descriptor rather than multiple descriptions (which we believed can cause confusion). Numerical descriptions were on a logarithmic scale. Descriptions of the range of probabilities in three of the 5 categories allow a degree of uncertainty to be specified.
2. Numerical with 10 points (N10pt) – Increased number of points on the scale allows us to address a point raised in the interviews, i.e. to allow very rare events to be represented. Since this scale is numerical only there is no opportunity to capture uncertainty.
3. Verbal numerical percentage (VN%) - In interviews experts expressed a preference for expressing probabilities as a percentage, hence the choice of this scale. However it does not allow for the specification of very small probabilities.
4. Direct statement of probability (DSP) – This scale allows the expert to state his/her beliefs without restriction, although it relies on the expert having training in probability theory
5. Direct statement of probability upper and lower limits (DSP_UL and DSP_LL) – This scale also relies on the expert having training in probability theory but

additionally allows a statement of levels of uncertainty, where wide values can show more uncertainty than tight values without any restriction.

3.2 Topics Used in Questions for which Probabilities are Assessed

The questions investigated embraced three scenarios and were deliberately selected to be understood by both company and student groups. Since the goal was to examine the effectiveness and ease-of-use of alternative probability methods, there was no constraint to focus on the questions about the types of the events that would be applicable in the real technical risk assessment. We also sought to ensure that we would be able to compare assessed subjective probabilities with the true probabilities of events; hence it was important to select topics for which observational data was, or would, be available over the course of the study. It was also important to include events whose probabilities of occurrence would span a meaningful range, especially in relation to rare events. We also needed to ensure that sufficient information was provided so that study respondents would be able to reason through their probability assessment.

One question focussed upon an upcoming event and asked "*State your subjective probability that, from all the countries competing in the 300 or so events 2012 Olympic Games in London, if an event is picked at random, what is the probability that China will win a medal?*".

A second question was on a measurable event and was phrased as follows "*A transit is the passage of a planet across the sun visible from Earth. Mercury is closest to the sun with the last transit occurring in 2006. What is the probability that there will be a transit of Mercury within the next two years?*".

The third question required probability to be assessed over a specific time period and was expressed as "*What is the probability that you will replace at least one tyre within the first 6 months of owning a new car? Assume your car is driven the annual mileage of 12,000miles (20,000km approx.) per year.*"

3.3 Design and Implementation of Questionnaire

The questionnaire was designed to ensure that all questions and all methods were included, albeit in a systematic arrangement designed to limit respondents referring to previously answered questions. Three questionnaire sets, each with a different ordering of questions, were prepared to control for any possible question/method ordering effects. Questionnaires from different sets were distributed at random to respondents in both company and student groups.

Two additional papers supplemented the questionnaire. A guidance document described the purpose of the study, an explanation of how to complete the questionnaire, including some examples questions/responses and directions for submission. We believed it was important to provide examples of how to approach probability elicitation to ensure all respondents were trained to at least a minimum level and also it is reported, by Clemen et al (2000) for example, that providing such guidance can improve accuracy and reduce the number of responses outside mathematical bounds. A feedback form was also designed to gather the impressions of respondents about their degree of understanding and comfort using each of the methods. Directly asking for feedback provided a way to obtain more information about the elicitation method and to provider views on their like or dislike of the methods tested. The feedback included an open question to uncover specific issues.

Given the scale and accessibility of respondents, the questionnaire was to be self-administered although different mechanisms were used for both, largely for reasons of practicality. Since the company group were situated in diverse locations across the UK, an electronic version of the questionnaire was distributed to 180 staff with responsibility for entering probability values in to the ERM system. Each of seven cohorts of the student group was given the questionnaire in paper form. Note that the seven cohorts were based in different geographical locations, although each cohort was effectively equivalent in its characteristics of the student group. One of the researchers (either second or third author) briefed the students on the purpose of the questionnaire and explained the guidance documentation. Completed questionnaires and feedback forms were returned anonymously to the researcher. Note that all questionnaires and feedback forms were anonymous. Response rates varied from 27% from the company to 70% for the students.

3.4 Findings from Analysis of Questionnaire Data

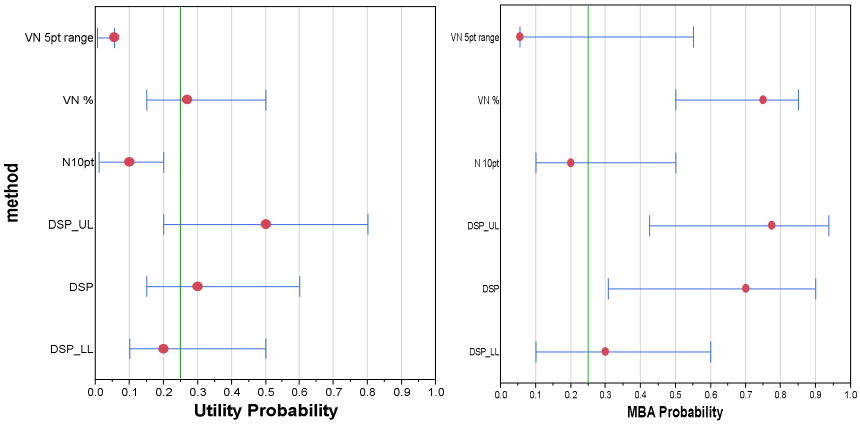
The three main outcomes emerging from analysis are discussed in turn.

3.4.1 Probability Scales and Anchoring

Our study finds that the two DSP methods (both point and interval values) perform least well for estimating the occurrence of events relative to the true probability. Both students and company groups tend to overestimate the chance of occurrence using the DSP method and generally seem unable to express low probability values using this method. This is in line with views expressed in existing literature. For example, according to Kahneman, (2011, p 324) "*people overestimate the probability of unlikely events*". So although the DSP methods offer most flexibility in specifying probability values, it is shown to be ineffective at eliciting low probabilities. For the low probability questions, the VN5pt range and the N10pt methods provided values closer to the true values than any of the other methods tested.

Figure 1 illustrates summary responses by both company and student groups to a single question (i.e. about the Olympics). Note that both groups anchor on the words

"expected" and "uncertain" for two of the methods (i.e. VN5pt and VN%), despite the fact that in each method the words correspond to very different probability values. This illustrates how people are drawn to qualitative statements and their own interpretation of what these words mean to them.



Company Students

Figure 1: Summary results for Olympics question showing verbal and numerical anchors with multiple scales for company and student groups

3.4.2 Differences between Company and Student Probability Assessments

From our analysis there is surprisingly little difference between the two groups in terms of their responses to all questions using all methods. A visual display synthesising all results is shown in Figure 2 and tends to show similar patterns between each group for all questions and methods. There are some exceptions. For example, for question 3 about the Olympics (i.e. the left of the figure) the response of the students is higher than that of the company group. Many of the students questioned were based in Asia and this perhaps might have influenced perceptions of the chance of China's medal winning chances?

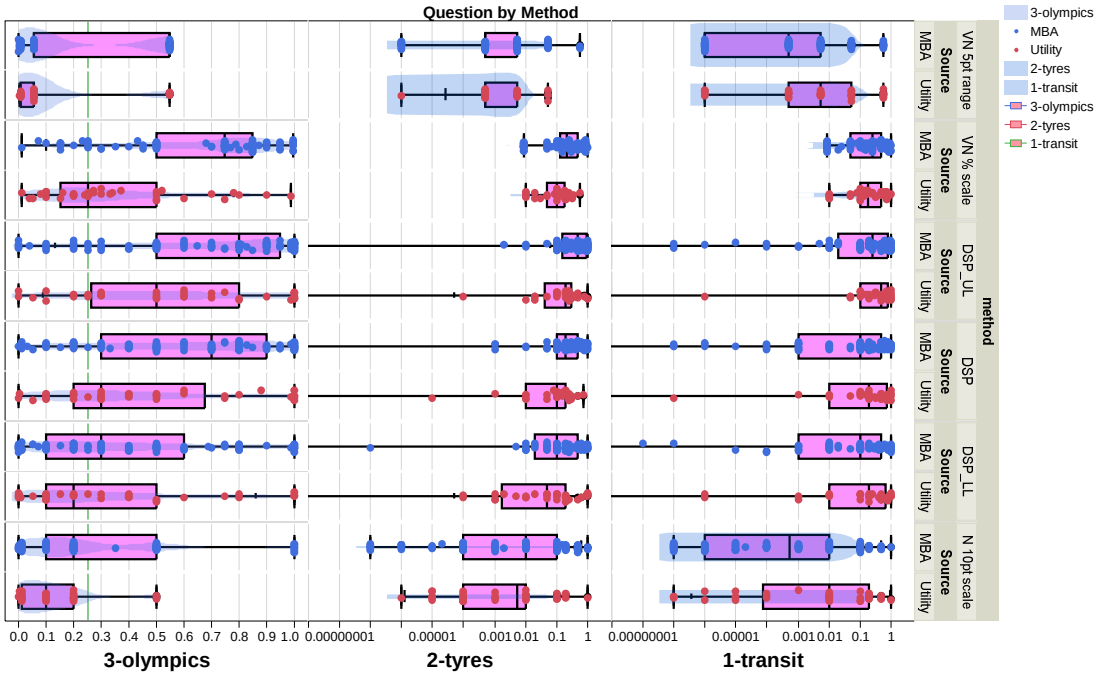


Figure 2: Summary results for each method (shown horizontally) each question (shown vertically) for company (in blue) and student (in red) groups showing a degree of similarity in response between two groups

Many elicitation studies, such Clemen et al (2000), do not explore the differences between the results drawn from students and other demographic groups, such as professional engineers. Naturally, each group has different skill sets, might have been exposed to statistical concepts at different times and might have different experiences and incentives. While we aimed to match our control group with those of engineering experts in company to some degree and to provide them with the same guidance, there were several major differences between them such as professional expertise and company culture. Hence we were surprised at the lack of any major systematic differences between the two groups.

3.4.3 Participant Feedback on Usability of Methods

Feedback from respondents suggested that the DSP methods were the most disliked with 39% (n=99) of students and 72% (n=37) of company group disliking it. The VN5pt method was the favourite for the company experts with 46% (n=23) stating this was their preferred method. This may be explained partially due to the similarity between their current method and the VN5pt Range method.

Qualitative feedback and analysis of quantitative results highlight that a preferred method is not necessarily the method that had the capability to provide better results, that is, closer to the true probability. For example, quantifying the mean square error between the judged probability and the true value for the Olympics question for all

methods except the DSP_LL and DSP_UL, for which such a computation is not meaningful, we find that the mean square error for the DSP method (i.e. 1.56) is around thirty time times larger than those for the VN%, VN5pt and N10pt methods (i.e. 0.058, 0.051, 0.045 respectively). Interestingly, the N10pt scale, which performs best, was originally developed to perform well for rare events and so it is reassuring that it also performs well for questions with probabilities that are not small.

4. Summary, Conclusions and Further Work

In this article we have discussed challenges affecting risk quantification in practice in relation to the elicitation of probabilities. There are many shortcomings of our study. For example, only a limited number of interviews were conducted in the case organisation, albeit with a representative selection of engineering experts with whom in-depth discussion was held. The experimental study was limited by the nature of the questions and the methods examined with data collected by alternative means for each of the company and student groups, although we aspire to control the design as far as was reasonably practicable. Notwithstanding these limitations, interesting insights have emerged that both extend and deepen understanding of issues discussed in the current literature.

Our interviews within the case organisation, for example, surfaced examples of bias, some of which can arguably be better controlled and hence their impact on risk assessments reduced through improved probability judgements. The importance of knowledge networks on the value of a probability elicitation process has emerged as an interesting insight. Acknowledging the nature of such networks and nurturing them to sustain or develop the knowledge base appears important in supporting experts in assessing uncertainty. We have shown that having a standard process and a management software tool in the form of an ERM system is not sufficient for supporting elicitation. The functionality and design features of the system require careful consideration to support probability elicitation. For example, screen layouts and the probability elicitation method used form a major part of the elicitation process, hence scientific principles of expert elicitation and probability theory should be integral to an ERM system. An ERM system will only support decisions that are as good as the data, judgemental or observed, that it contains. Any shortcomings in the elicitation of probability will influence the quality and validity of information provided by an ERM system.

The experimental study of alternative methods has shown that asking for a direct statement of probability (i.e. DSP) is least effective in providing an accurate assessment, especially for very low probability values, even though it is the method that allows the expert most control and flexibility in the specification of a probability value. Interestingly all study respondents across the company and student groups appear less able to assess low probability values. Methods that use of a predefined numerical scale, either with or without words yielded lower probability estimates than were observed for the selected questions we asked. The company group claimed

more confidence using the VN5pt method than the students, but this is not surprising given this is close to the existing organisational standard method, and there was evidence of anchoring on the numerical values rather than the verbal cues. Scales with verbal anchors appear to cue respondents to select a particular probability value and can affect the choice of probability value. This appears particularly important where there are higher levels of uncertainty, or unfamiliarity with the method, thus highlighting the importance of on-going training and feedback to experts in the use of the probability elicitation method selected. All respondents prefer methods with a scale, although the preferred method for both groups was not necessarily the one that provided the most accurate results. It seems obvious to state, but we cannot underestimate the importance of selecting a probability elicitation method that fits the application context and the range of probabilities being assessed.

Future work could investigate additional VN descriptors to establish their impact on the bias occurring and to further understand the nature of anchoring to a particular value. For example, a magnifier scale (e.g. Gurmankin et al, 2005) which involves great articulation of one part of the scale could be considered but we might expect that this would require extensive training for proper use.

5. References

1. Health and Safety Executive (2001). "Reducing Risk, Protecting People." Discussion Document. HSE DDE11 C 150: 5.
2. O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Gathwaite, P. H., Jenkinson, D. J., Oakley, J. E. & Rakow, T. (2006), "Uncertain judgements. Eliciting experts' probabilities", Statistics in Practice, John Wiley & Sons, Chichester.
3. Cooke, Roger M. "Experts in Uncertainty. Opinion and Subjective Probability in Science". Environmental Ethics and Science Policy Series. New York: Oxford University Press, 1991.
4. Kahneman, D., P. Slovic, And Tversyky, A. (1982). "Judgement Under Uncertainty: Heuristics and Biases". Cambridge, Cambridge University Press.
5. Van Der Gaag ,L.C., Renooij ,S., Witteman C.L.M., Aleman , B. M. P. Taal, B. G. "How to Elicit Many Probabilities" (1999), Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence
6. Renooij, S. (2001), "Probability elicitation for belief networks: issues to consider". The Knowledge Engineering Review, 16(3), 255-269
7. Clemen, R. T., G. W. Fischer, and Winkler, R.L. (2000). "Assessing Dependence: Some Experimental Results." Management Science 46(8): 1100-1115.
8. Kahneman, D. 2011 "Thinking Fast and Slow", Penguin, Great Britain
9. Gurmankin, A.D., Helweg-Larsen, M., Armstrong, K., Kimmel, S.E., Volpp, K.G.M., 2005, "Comparing the Standard Rating Scale and the Magnifier Scale for Assessing Risk Perceptions", Medical Decision Making, 25: 560

6. Acknowledgements

This work would not have been possible without the participation of the MBA students and the engineers within the utility who kindly participated in the study. We are also grateful to risk analysis staff in the case organisation for their assistance in identifying experts and feedback throughout the project.

On Combined Data Under Competing Risks

Tahani Coolen-Maturi[*,1] **and Frank P.A. Coolen**[+]

[*]Durham University Business School, Durham University, UK.
[+]Department of Mathematical Sciences, Durham University, UK.

Abstract

Maturi et al. [14] presented the nonparametric predictive inference (NPI) approach for competing risks data, in particular addressing the question due to which of the competing risks the next unit will fail. Recently, Coolen-Maturi and Coolen [8] considered further aspects which are closely related to that paper, in particular the effects of unobserved, re-defined, unknown or removed competing risks. In this paper, we introduce how the NPI approach can be used to deal with situations where units are not all at risk from all competing risks. This may typically occur if one combines information from multiple samples, which can e.g. be related to further aspects of units that define the samples or groups to which the units belong or to different applications where the circumstances under which the units operate can vary. We study the effect of combining the additional information from these multiple samples, so effectively borrowing information on specific competing risks from other units, on the inferences.

1. Introduction

Nonparametric predictive inference (NPI) is a statistical method based on Hill's assumption $A_{(n)}$ [11], which gives a direct conditional probability for a future observable random quantity, conditional on observed values of related random quantities [1, 2]. $A_{(n)}$ does not assume anything else, and can be interpreted as a post-data assumption related to exchangeability [10]. Inferences based on $A_{(n)}$ are predictive and nonparametric, and can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use such information, e.g. to study effects of additional assumptions underlying other statistical methods. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides bounds for probabilities via the 'fundamental theorem of probability' [10]. These bounds are lower and upper probabilities in imprecise probability theory [1, 15, 16]. Short introductions to NPI, imprecise probability and its use in reliability have recently been presented in [3], [4] and [6], respectively.

In reliability and survival analysis, data on event times are often affected by right-censoring, where for a specific unit or individual it is only known that the event has not yet taken place at a specific time. Coolen and Yan [7] presented a

[1]Corresponding author: tahani.maturi@durham.ac.uk

generalization of $A_{(n)}$, called rc-$A_{(n)}$, which is suitable for right-censored data. In comparison to $A_{(n)}$, rc-$A_{(n)}$ uses the additional assumption that, at the moment of censoring, the residual lifetime of a right-censored unit is exchangeable with the residual lifetimes of all other units that have not yet failed or been censored. The assumption rc-$A_{(n)}$ underlies the inferences in this paper, for more details we refer to [7, 17].

Coolen et al. [5] introduced NPI for some reliability applications, including upper and lower survival functions for the next future observation, illustrated with an application with competing risks data. They illustrated the lower and upper marginal survival functions, so each restricted to a single failure mode. Maturi et al. [14] presented NPI for competing risks data, in particular addressing the question due to which of the competing risks the next unit will fail. Related to this approach, Coolen-Maturi and Coolen [8] considered the effects of unobserved, re-defined, unknown or removed competing risks.

In NPI for competing risks [14], it is assumed that there are K failure modes and a unit fails due to the first occurrence of a failure mode, which is identified with certainty. We should point out that, in this paper, we will use the terms 'failure mode' and 'competing risk' interchangeably with the same meanings. Let X_{n+1} denote the failure time of a future unit, based on n observations, and let the corresponding notation for the failure time including indication of the actual failure mode k be $X_{k,n+1}$. It is important to emphasize that $X_{k,n+1}$ is interpreted as the random failure time of a future unit which is only at risk from failure mode k. Different failure modes are assumed to occur independently. The competing risk data per failure mode consist of a number of observed times of failures caused by the specific failure mode considered, and right-censoring times caused by other failure modes or other reasons for right-censoring. Hence rc-$A_{(n)}$ can be applied per failure mode k for inference on $X_{k,n+1}$.

Suppose that, in the available data, u_k failures are caused by failure mode k, at times $x_{k,1} < x_{k,2} < \ldots < x_{k,u_k}$, and let $n - u_k$ be the number of the right-censored observations, $c_{k,1} < c_{k,2} < \ldots < c_{k,n-u_k}$, corresponding to failure mode k; these may be failure times due to other (independent) failure modes, or observations that are right-censored for other reasons, where it is assumed throughout that such censoring processes are independent of $X_{k,n+1}$. For notational convenience, let $x_{k,0} = 0$ and $x_{k,u_k+1} = \infty$. Suppose further that there are s_{k,i_k} right-censored observations in the interval (x_{k,i_k}, x_{k,i_k+1}), denoted by $c_{k,1}^{i_k} < c_{k,2}^{i_k} < \ldots < c_{k,s_{k,i_k}}^{i_k}$, so $\sum_{i_k=0}^{u_k} s_{k,i_k} = n - u_k$. It should be emphasized that we do not assume that each unit considered must actually fail, if a unit does not fail then there will be a right-censored observation recorded for this unit for each failure mode, as we assume that the unit will then be withdrawn from the study, or the study ends, at some point. The random quantity representing the failure time of the next unit, with all K failure modes considered, is $X_{n+1} = \min_{1 \leq k \leq K} X_{k,n+1}$.

For $i_k = 0, 1, \ldots, u_k$, let $t_{k,i_k^*}^{i_k} = c_{k,i_k^*}^{i_k}$ (i.e. censoring time) for $i_k^* = 1, 2, \ldots, s_{k,i_k}$ and $t_{k,i_k^*}^{i_k} = x_{k,i_k}$ (i.e. failure time or time 0) for $i_k^* = 0$. For notational convenience, let $t_{k,s_{k,i_k}+1}^{i_k} = t_{k,0}^{i_k+1} = x_{k,i_k+1}$ for $i_k = 0, 1, \ldots, u_k - 1$. Let $\tilde{n}_{c_{k,r}}$ and $\tilde{n}_{t_{k,i_k^*}^{i_k}}$ be the

number of units in the risk set just prior to time $c_{k,r}$ and $t^{i_k}_{k,i^*_k}$, respectively, with the definition $\tilde{n}_0 = n+1$ for ease of notation. The risk set at a certain time contains all units that have not failed or been right-censored before that time, and hence are indeed still at risk. The NPI lower and upper survival functions for the failure time of the next unit due to failure mode k, so if the unit were only at risk from this failure mode, are denoted by $\underline{S}_{X_{k,n+1}}(t)$ and $\overline{S}_{X_{k,n+1}}(t)$, respectively, and are as follows [5, 14]. For $t \in (t^{i_k}_{k,a_k}, t^{i_k}_{k,a_k+1}]$ with $i_k = 0, 1, \ldots, u_k$ and $a_k = 0, 1, \ldots, s_{k,i_k}$,

$$\underline{S}_{X_{k,n+1}}(t) = \frac{1}{n+1}\, \tilde{n}_{t^{i_k}_{k,a_k+1}} \prod_{\{r:c_{k,r}<t^{i_k}_{k,a_k+1}\}} \frac{\tilde{n}_{c_{k,r}}+1}{\tilde{n}_{c_{k,r}}} \tag{1}$$

and for $t \in [x_{k,i_k}, x_{k,i_k+1})$ with $i_k = 0, 1, \ldots, u_k$,

$$\overline{S}_{X_{k,n+1}}(t) = \frac{1}{n+1}\, \tilde{n}_{x_{k,i_k}} \prod_{\{r:c_{k,r}<x_{k,i_k}\}} \frac{\tilde{n}_{c_{k,r}}+1}{\tilde{n}_{c_{k,r}}} \tag{2}$$

While predictive inference, as considered in this approach, is different to estimation, as it explicitly considers a single future unit instead of estimating characteristics of a population distribution, it is interesting to mention that these NPI lower and upper survival functions bound the well-known Kaplan-Meier estimator [12], which is the nonparametric maximum likelihood estimator for the population survival function in case of lifetime data with right-censored observations, for more details we refer to [7, 9].

If all the units are censored with regard to failure mode k (e.g. if all units failed due other failure modes), then the lower and upper survival functions in (1) and (2) are equal to [8],

$$\underline{S}_{X_{n+1}}(t) = \frac{\tilde{n}_{t^{i_k}_{k,a_k+1}}}{\tilde{n}_{t^{i_k}_{k,a_k+1}}+1} \qquad \text{and} \qquad \overline{S}_{X_{n+1}}(t) = 1 \tag{3}$$

If the next unit considered is at risk from K independent failure modes, so with its failure time given by $X_{n+1} = \min_{1\le k\le K} X_{k,n+1}$, then the NPI lower and upper survival functions for its failure time are denoted by $\underline{S}_{X_{n+1}}(t)$ and $\overline{S}_{X_{n+1}}(t)$, respectively, and are equal to

$$\underline{S}_{X_{n+1}}(t) = \prod_{k=1}^{K} \underline{S}_{X_{k,n+1}}(t) \qquad \text{and} \qquad \overline{S}_{X_{n+1}}(t) = \prod_{k=1}^{K} \overline{S}_{X_{k,n+1}}(t) \tag{4}$$

2. NPI for Combined Data Under Competing Risks

We now present a generalization of the NPI approach to competing risks, by considering the important situation of different groups of units, such that units from the same group are at risk from the same set of competing risks, but these sets differ for the different groups. Of course, it is typically assumed that there

is at least some overlap between the sets of competing risks for different groups. In this case, the information in data from different groups about a specific failure mode, that applied to these groups, can be used to enhance inferences for a unit at risk from this failure mode. To enable such learning from information about other groups, we make the important assumption, throughout this paper, that a failure mode affects all units that are at risk from it in the same way, no matter which group the unit belongs to. And, as mentioned before but crucial to the approach, we assume throughout this paper that all failure modes that affect a unit do so independently. This section starts with an introduction of further notation and detailed discussion of the scenario considered in this paper in Section 2.1. In Section 2.2 we consider inference about a specific failure mode, using data from all groups of units that were at risk from this failure mode. Section 2.3 combines such inferences to NPI lower and upper survival functions for a future unit from a particular group, so taking all failure modes that affect such a unit into account. An illustrative example is presented in Section 3.

2.1. Notation and setting

In addition to notation introduced above, suppose we have M groups to which individual units can belong, denoted by $G_1, \ldots, G_m, \ldots, G_M$. The sets of failure modes that affect units are different per group. In medical survival analysis, such groups can be defined, for example, by covariates indicating sex or aspects of lifestyle. In reliability analyses, one can think about units that are used in different production processes or at different sites. We suppose that we have failure time data from each group, with sample size $n_m > 0$ for group $m \in \{1, \ldots, M\}$. We assume that in total there are K competing risks, denoted by $R_1, \ldots, R_K$. Let d_{mk} be an indicator function defined as

$$d_{mk} = \begin{cases} 1 & \text{if units in group } G_m \text{ can fail due to risk } R_k \\ 0 & \text{if not} \end{cases}$$

We should emphasize that for $d_{mk} = 1$ we may actually have observed failures in group G_m due to risk R_k or this may not be the case; we assume in this paper that the values d_{mk} are known with certainty. With regard to the data, let

$$d^*_{mk} = \begin{cases} 1 & \text{if at least one failure due to } R_k \text{ has been observed for group } G_m \\ 0 & \text{if no failure due to } R_k \text{ has been observed for group } G_m \end{cases}$$

So $d_{mk} = 0$ logically implies $d^*_{mk} = 0$, but if $d_{mk} = 1$ the corresponding d^*_{mk} can be either 1 or 0.

For an index set $J \subseteq \{1, 2, \ldots, K\}$ we will be interested in the NPI lower and upper survival functions for units at risk from all failure modes R_k for $k \in J$ (and no other failure modes); these are denoted by $\underline{S}_J$ and $\overline{S}_J$. Let J_m be such an index set referring to the set of failure modes due to which units in group G_m can fail, so $J_m = \{k : d_{mk} = 1\}$. We further define the index set of observed risks in the data for group G_m by $J^*_m = \{k : d_{mk} = 1 \text{ and } d^*_{mk} = 1\}$.

Similarly, we define the index set of groups whose units can fail due to failure mode R_k as $I_k = \{m : d_{mk} = 1\}$, so $I_k \subseteq \{1, 2, \ldots, M\}$. We will consider the

information about failure mode R_k in the data sets for the groups G_m with $m \in I_k$. Based on these combined data, we consider inference on a future unit under the assumption that it is only at risk from failure mode R_k; to emphasize that such an inference will be based on the data from all groups G_m with $m \in I_k$, we will denote the NPI lower and upper survival functions for such a future unit by $\underline{S}_{I_k}$ and $\overline{S}_{I_k}$. This notation can again be extended to indicate if we only take information into account from groups for which this specific failure mode R_k has actually been observed; we then denote the index set corresponding to the groups for which this failure mode has been observed by $I_k^* = \{m : d_{mk} = 1 \text{ and } d_{mk}^* = 1\}$, with similar extension of the notation for the corresponding NPI lower and upper survival functions.

2.2. Failure mode R_k

It is interesting to consider inference about a failure mode R_k, that is inference about the failure time of a future unit which is only at risk from failure mode R_k. This is of interest in its own right, to learn about this failure mode, but also for its use in the competing risks scenario for a future unit which is at risk from several failure modes, which will be presented in Section 2.3. We use all data from all the groups whose units could have failed due to this failure mode, so groups G_m with $m \in I_k = \{m : d_{mk} = 1\}$. With interest only in failures due to R_k, all failure times in the data for these groups that were caused by other failure modes are considered as right-censored observations. Note that it is irrelevant here, due to the assumptions of independent failure modes, that units from different groups which are included in the data for this inference on R_k will not all have been at risk from the same failure modes; it is of no relevance which specific other failure modes caused the failures at times which are treated here as right-censored observations. The lower and upper survival functions for a future unit which is only at risk from failure mode R_k, and based on all data from groups G_m with $m \in I_k$, are (as mentioned in Section 2.1) denoted by $\underline{S}_{I_k}$ and $\overline{S}_{I_k}$, these are straightforwardly derived from (1) and (2), respectively, using all the data from groups G_m with $m \in I_k$ as explained here. This predictive inference is based on information from $\sum_{m=1}^{M} n_m \times d_{mk}$ units in the data set.

If we consider only the observed failure modes, when interested in failures due to R_k, we use the data from all groups in which at least one unit has failed due to this failure mode, so groups G_m with $m \in I_k^* = \{m : d_{mk} = 1 \text{ and } d_{mk}^* = 1\}$. The NPI lower and upper survival functions for a future unit which is only at risk from failure mode R_k, and based on data from groups G_m with $m \in I_k^*$, are denoted by $\underline{S}_{I_k^*}$ and $\overline{S}_{I_k^*}$; these are straightforwardly derived from (1) and (2). This predictive inference is based on information from $\sum_{m=1}^{M} n_m \times d_{mk} \times d_{mk}^*$ units in the data set.

2.3. Unit from group G_m

We now consider inference about the failure time of the next unit from a specific group G_m, which can fail due to all failure modes that can affect units from this

group, so failure modes R_k with $d_{mk} = 1$, hence with $k \in J_m$. The NPI lower and upper survival functions for the failure time of this future unit from group G_m are given by

$$\underline{S}_{J_m}(t) = \prod_{k \in J_m} \underline{S}_{I_k}(t) \quad \text{and} \quad \overline{S}_{J_m}(t) = \prod_{k \in J_m} \overline{S}_{I_k}(t) \tag{5}$$

where $\underline{S}_{I_k}$ and $\overline{S}_{I_k}$ are as presented in Section 2.2. Hence, this inference combines, for each failure mode that is being considered, all the data from different groups as described in Section 2.2.

We can also consider inference about the next unit from group G_m but using only data from groups for which the relevant failure modes actually have been observed. Then the NPI lower and upper survival functions are

$$\underline{S}_{J_m^*}(t) = \prod_{k \in J_m^*} \underline{S}_{I_k^*}(t) \quad \text{and} \quad \overline{S}_{J_m^*}(t) = \prod_{k \in J_m^*} \overline{S}_{I_k^*}(t) \tag{6}$$

where $\underline{S}_{I_k^*}$ and $\overline{S}_{I_k^*}$ are as presented in Section 2.2.

This inference for the failure time of a future unit from group G_m is the main novelty presented in this paper. Such use of information from other groups is also known as 'borrowing information' from other groups in order to derive stronger inferences. Note that such borrowing of information from other groups is done separately for each relevant failure mode, and that the assumption that different risks affect the units independently is required to combine the NPI lower and upper survival functions for the different failure modes in this way.

3. An illustrative example

In this example, a well-known data set from the literature [13] is used to illustrate the NPI method proposed in this paper. The data contain information about 36 units of a new model of a small electrical appliance which were tested, and where the lifetime observation per unit consists of the number of completed cycles of use until the unit failed (we interpret this number as a continuous quantity). To illustrate our method, we have divided this data set into three groups, G_1, G_2 and G_3, as presented in Table 1, which also includes the specific failure mode (R) that caused the unit to fail. In the study, there were 18 different ways in which an appliance could fail, so 18 failure modes, but only 7 of them have been observed. Failure modes R_9 and R_6 cause the largest number of units to fail, 17 units and 7 units, respectively, while failure modes R_2, R_5, R_{10} and R_{15} each cause two units to fail, and failure mode R_1 causes one unit to fail. Three units in the test did not fail before the end of the experiment, so for these units we have right-censored observations (2565, 6367 and 13403) for all failure modes considered, indicated by '0' for the failure mode in Table 1. With this grouping of the data, there are 4 observed failure modes per group: failure modes R_1, R_6, R_9 and R_{10} have been observed in group G_1, failure modes R_2, R_9, R_{10} and R_{15} have been observed in group G_2, and failure modes R_5, R_6, R_9 and R_{15} have been observed in group G_3.

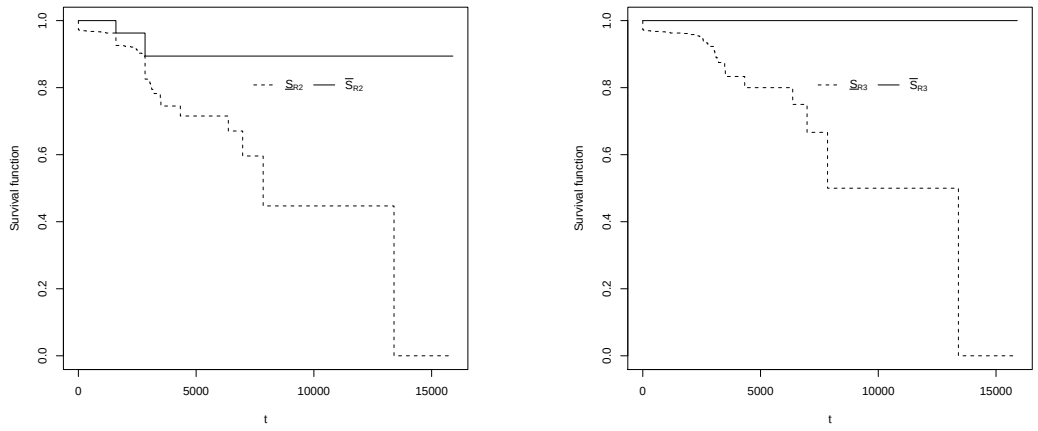
G_1		G_2		G_3	
# cycles	R	# cycles	R	# cycles	R
11	1	35	15	49	15
381	6	958	10	170	6
708	6	1167	9	329	6
1925	9	1594	2	1062	5
2223	9	1990	9	2400	9
2327	6	2471	9	2451	5
2702	10	2551	9	2761	6
3035	9	2565	0	3034	9
3059	6	2568	9	3112	9
3504	9	2831	2	3478	9
6976	9	3214	9	6367	0
7846	9	4329	9	13403	0

Table 1: Failure data for electrical appliance test

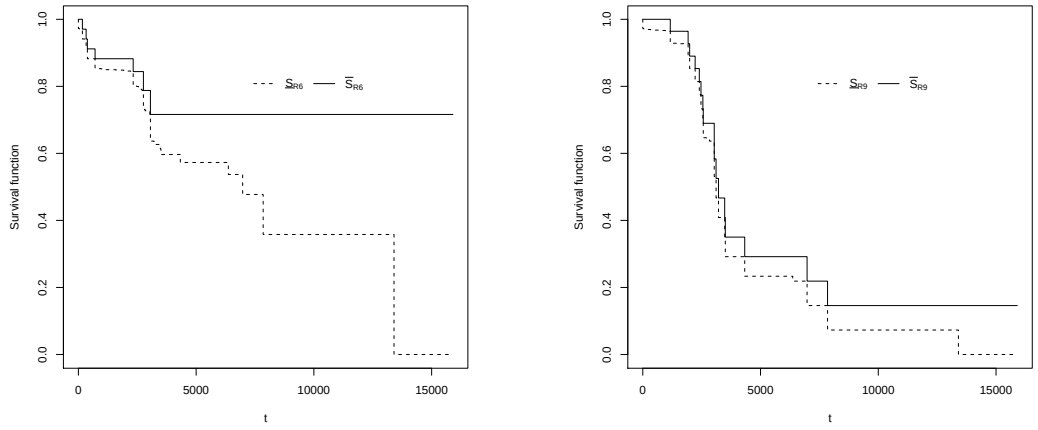
We start by considering inference about a specific failure mode R_k, as described in Section 2.2. We discuss the following two options. First, we assume that the units in all groups could have failed due R_k, so we use the data for all units. Secondly, we assume that only units in groups in which R_k has actually been observed, were at risk from R_k, so only data from such groups are included for the inferences. This inference is in terms of the failure time of a future unit that is at risk from R_k only, and based on the data according to the specific assumption under these two possible scenarios.

Figure 1 presents the NPI lower and upper survival functions for the next unit which is at risk from R_k only, for $k = 2, 3, 6, 9$, under the assumption that all units in the three groups of the data set had been at risk from R_k, so based on all 36 observations. In this figure we denote the NPI lower and upper survival functions by $\underline{S}_{R_k}$ and $\overline{S}_{R_k}$, $k = 2, 3, 6, 9$, respectively. Note that we only observed, in total, two failures due to failure mode R_2, and we did not observe any failure due to failure mode R_3. This is reflected in Figures 1(a) and 1(b), respectively, as the NPI upper survival function only decreases at an observed failure time caused by the specific R_k, while the NPI lower survival function decreases at every observation, so both at failure times and right-censoring times with regard to R_k, the latter being failure times caused by other failure modes as well as actually right-censored observations. The NPI upper survival functions remain quite large for Figures 1(a)-1(c), of course particularly in Figure 1(b) where it remains at value 1. This reflects that these data provide little evidence (or none at all for R_3) against the possibility that units may actually not fail due to the failure mode that is being considered. The corresponding NPI lower survival function reflects the evidence in the data in favour of survival past time t, which decreases at every observation because the number of items in the data that were at risk at time t decreases. Beyond the largest observation, $t = 13403$, the NPI lower survival function is equal to zero, reflecting that the data do not contain evidence in favour of further survival.

We now consider the same inference, so failure time of a future unit which is assumed to be only at risk from failure mode R_k, but we assume explicitly



(a) R_2 (b) R_3

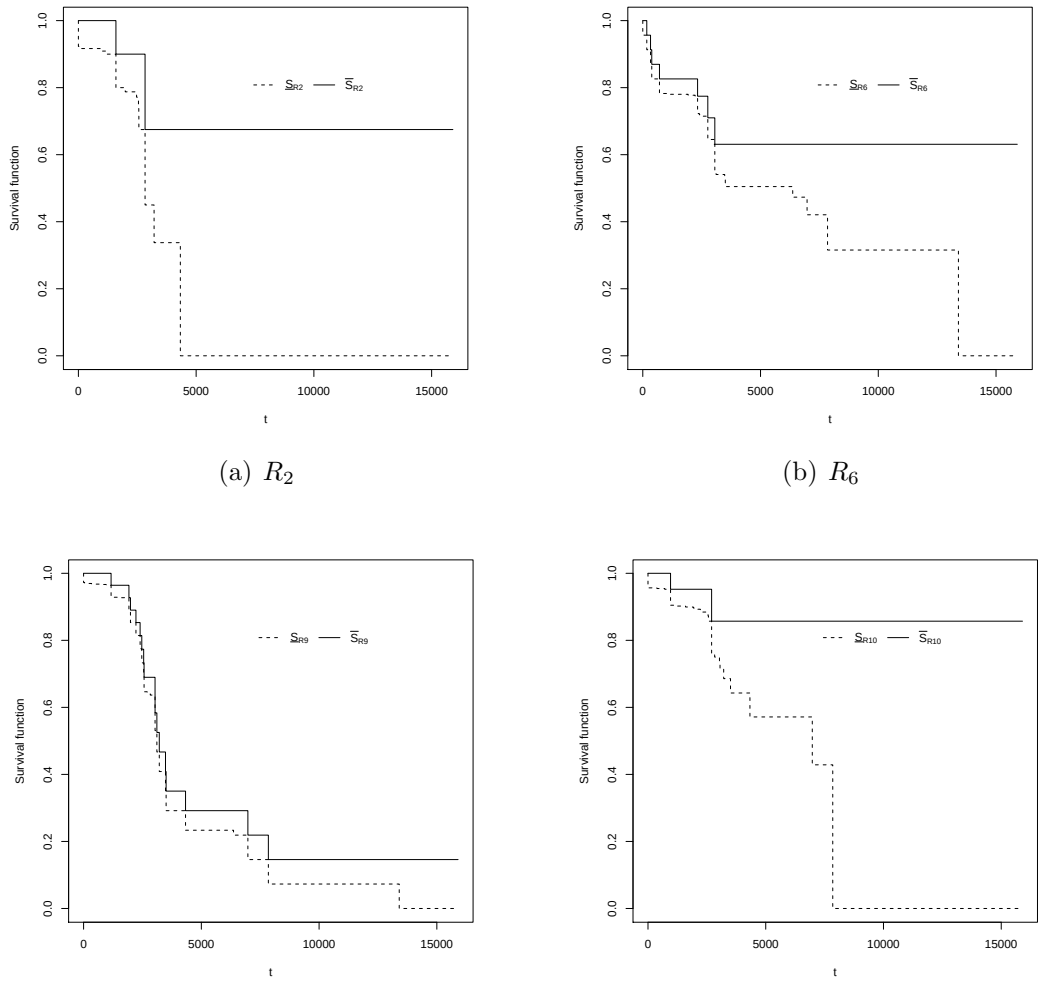


(c) R_6 (d) R_9

Figure 1: Lower and upper survival functions for R_k, $k = 2, 3, 6, 9$ (data: all groups)

that this failure mode only affected units in the group(s) in which it was actually observed. It should be emphasized that the decision on whether this is appropriate use of the data, or data from all groups can be used as presented above, is up to the topic expert and necessarily based on detailed information about the actual setting, clearly the data do not provide information to distinguish between these two scenarios. One may have an intermediate case where it was known that the failure mode did affect units in some groups where it was not observed, but not all such groups. We do not discuss this further in this paper, the methodology can straightforwardly be generalized to allow this.

For the data in this example, separated in three groups as given in Table 1, the NPI lower and upper survival functions for some failure modes are presented in Figure 2, under the assumption that a specific failure mode only affected units in groups where it has actually been observed. This implies that inference for failure



(a) R_2 (b) R_6

(c) R_9 (d) R_{10}

Figure 2: Lower and upper survival functions for R_k, $k = 2, 6, 9, 10$ (data: only groups for which R_k has been observed)

mode R_9 is based on data from all the groups, as this failure mode was observed to cause failures in each group. Hence, the NPI lower and upper survival functions for R_9 in this case, as presented in Figure 2(c), are identical to those presented in Figure 1(d), because both cases take the observations from all groups into account. As another extreme situation, for any non-observed failure mode there is now no meaningful inference, as no data are available to base inferences on. For example, using the data from all groups, we did get a non-trivial inference for R_3, due to the fact that all items had functioned for some time without failing due to this failure mode. But no such information is available under the assumption considered now. In the first situation, the information was reflected through the NPI lower survival function in Figure 1(b), which decreased at each right-censored observation. In this case with no data we could only provide the vacuous inference of NPI lower survival function equal to 0 for all $t > 0$. Note that, based on no

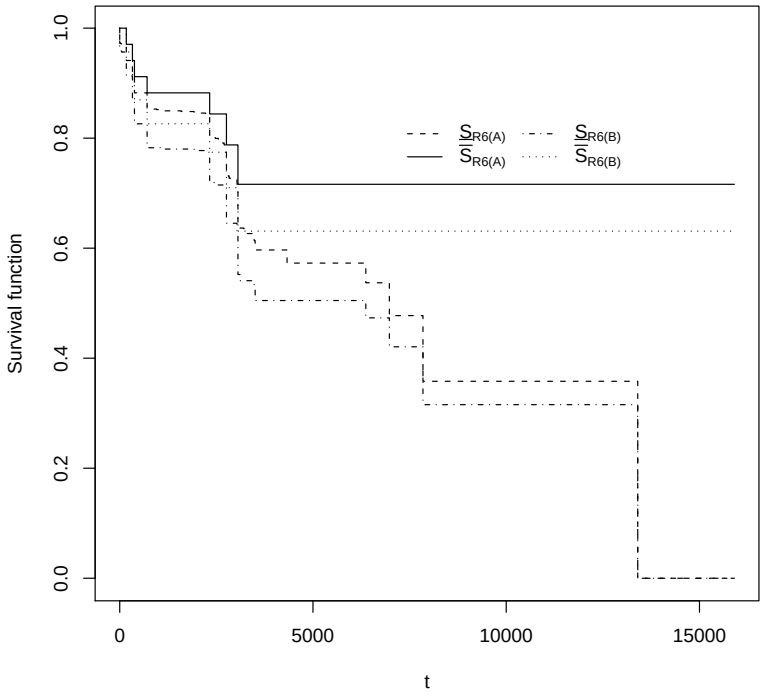


Figure 3: Lower and upper survival functions for R_6: (A) using all data, (B) only data from G_1 and G_3.

data information, we could only conclude an NPI upper survival function equal to 1 in this case, which is identical to the one for the situation above, which indeed reflects that there was no strong information in the data that was actually against the possibility that this failure mode might never lead to failures.

Figure 2 also presents the NPI lower and upper survival functions for failure mode R_2, using data from group G_2 only, failure mode R_6, using data from groups G_1 and G_3, and failure mode R_{10}, using data from groups G_1 and G_2. Of course, similar plots could be presented for the NPI lower and upper survival functions for the other failure modes. It is interesting to compare the NPI lower and upper survival functions for R_6 in Figure 1(c) and Figure 2(b), as these are based on different information. To emphasize the differences, these functions are presented together in Figure 3. The two NPI upper survival functions decrease only at the 7 observed failure times due to R_6. However, using all data (indicated by (A) in Figure 3) implies that more units did not fail due to R_6, hence the data contain less evidence against survival at any time t past the first time of a failure due to R_6 than for the situation where only data from groups G_1 and G_3 is used (indicated by (B) in Figure 3). This is reflected by the fact that the NPI upper survival function for the first situation is greater than for the second situation, beyond the first failure time caused by R_6 (up to that time both are equal to 1). The NPI lower survival functions for R_6 in these two scenarios differ more, due to the fact

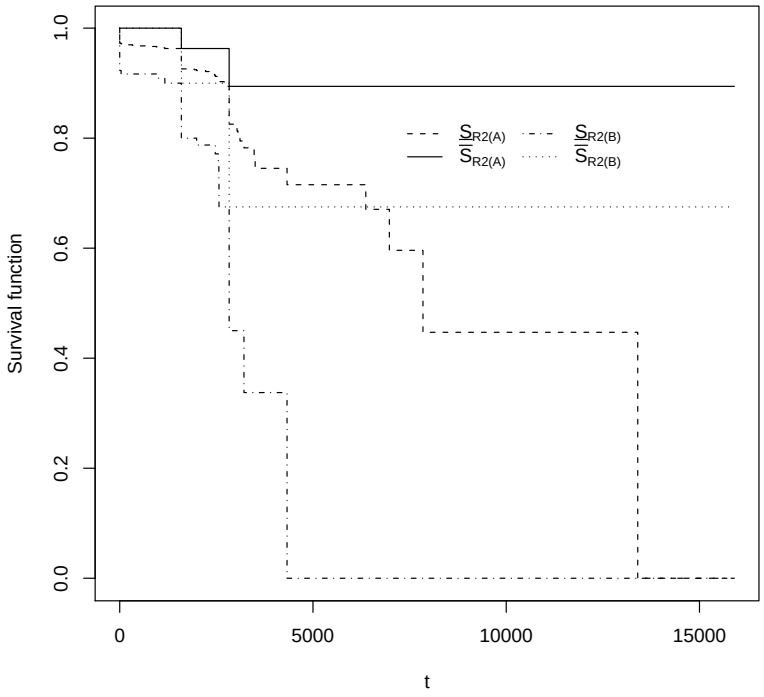


Figure 4: Lower and upper survival functions for R_2: (A) using all data, (B) only data from G_2.

that these functions decrease at every observation in the data set, so also at right-censored observations, and the use of the data from G_2 in the first case discussed above (indicated by (A) in Figure 3) but not in the second case (indicated by (B) in Figure 3) implies that with more data used the latter lower survival function decreases at more time points. The lower survival function when all data are used here is greater than the one based on data from G_1 and G_3 only, for all t up to the largest observation. This is because the additional information only consists of right-censoring times and hence provides some more information in favour of survival at any time until the largest observation, after which both these lower survival functions are equal to 0, reflecting that the data do not contain strong information in favour of surviving times beyond the largest observation.

The same properties hold for the NPI lower and upper survival functions for failure mode R_2 in Figure 1(a) and Figure 2(a), with the former based on the observations from all groups, so from 36 units, but the latter only on the data from the 12 units in group G_2. These lower and upper survival functions are also presented together in Figure 4, to show the differences more clearly. In this latter case (indicated by (B) in Figure 4), the NPI lower survival function decreases only at 12 time points, the last one at $t = 4329$ from which moment on the lower survival function is equal to 0. With the largest observation in all combined data being at $t = 13403$, the lower survival function based on all combined data

(indicated by (A) in Figure 4) only becomes 0 at that time point, so there is a substantial difference between these two lower survival functions. While the two NPI upper survival functions for these cases both decrease only at the two observed failure times due to R_2, the inclusion of many more right-censored observations (with regard to this failure mode) in the first case leads to a substantial difference between these two upper survival functions.

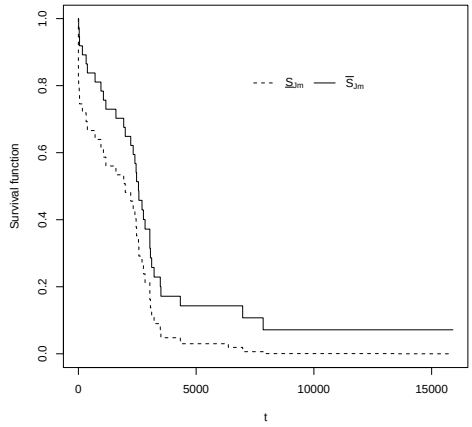
We now use this example to illustrate the methods presented in Section 2.3, so we consider the failure time of the next unit from group G_m, where again assumptions are required about which failure modes are assumed to possibly affect this unit. There are several possible assumptions, we present the ones which we consider to be of most interest.

First, we assume that all failure modes that have been observed at least once affected the units from all three groups, so these are R_k for $k = \{1, 2, 5, 6, 9, 10, 15\}$. We are interested in the failure time of a future unit which is at risk from precisely these 7 failure modes, so this could be a unit from any of the groups G_1, G_2 or G_3. The NPI lower and upper survival functions for such a future unit from any group, in this case, are presented in Figure 5(a), and they are derived by (for all $m = 1, 2, 3$)

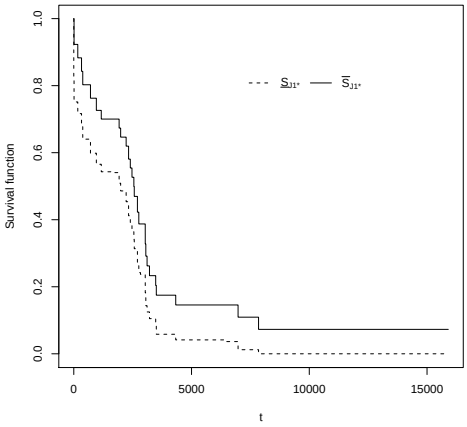
$$\underline{S}_{J_m}(t) = \prod_{k \in \{1,2,5,6,9,10,15\}} \underline{S}_{I_k}(t) \quad \text{and} \quad \overline{S}_{J_m}(t) = \prod_{k \in \{1,2,5,6,9,10,15\}} \overline{S}_{I_k}(t)$$

It should be emphasized that here we combine the NPI lower and upper survival functions for the individual failure modes that can affect the unit of interest, while we have used all available data to first derive these lower and upper survival functions along the lines as presented in Section 2.2 and illustrated earlier in this example. We could have added one or more of the identified but unobserved failure modes to this approach, if we wished to assume that these could indeed also affect such a future unit. Due to the NPI upper survival functions for such a failure mode being equal to 1 for all t, the upper survival function $\overline{S}_{J_m}$ would not be affected. However, the lower survival function $\underline{S}_{J_m}$ would be multiplied by the lower survival function(s) for such additional failure mode(s), which under the assumptions that they could have affected all data observations would (all) be equal to the lower survival function for R_3 in Figure 1(b). This latter lower survival function decreases at the same 36 time points as $\underline{S}_{J_m}$, namely all observation times in the data, and hence the resulting lower survival function for the next unit would be less than the $\underline{S}_{J_m}$ given in Figure 5(a). This would show the effect of the unit being possibly affected by more failure modes (leading to the decrease of the lower survival function), but not necessarily so as these failure modes have not yet been observed so there is no strong evidence that they will actually have an effect (shown by the unchanged upper survival function).

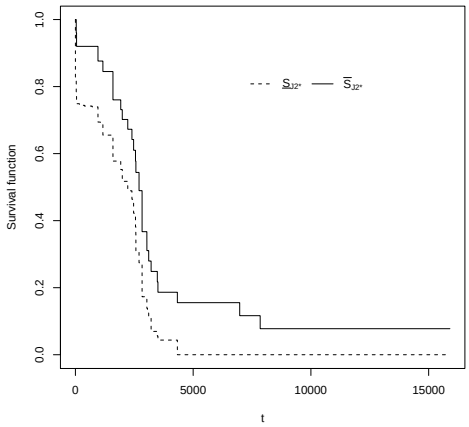
As a second scenario of interest, we assume that a future unit of group G_m is only affected by failure modes already observed for that group. In addition, we assume that, for as far as the data are concerned, only units in groups for which a particular failure mode has been observed were actually affected by it and hence only data from these groups are used for the inference about a specific failure



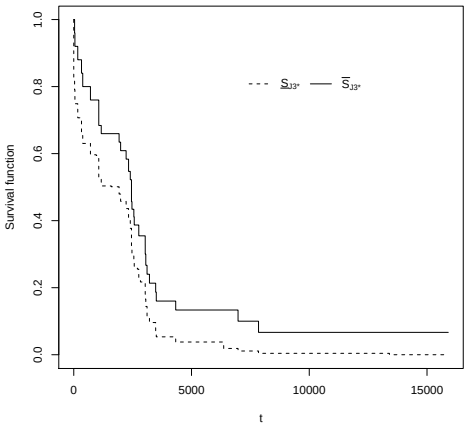
(a) G_m, $m = 1, 2, 3$, using all data



(b) G_1, using data where failure modes are observed



(c) G_2, using data where failure modes are observed



(d) G_3, using data where failure modes are observed

Figure 5: Lower and upper survival functions for G_m, $m = 1, 2, 3$

mode; this means using the NPI lower and survival functions shown in Figure 2 (and the corresponding ones for failure modes not shown in that figure). The resulting NPI lower and upper survival functions for units from groups G_1, G_2 and G_3 are presented in Figures 5(b), 5(c) and 5(d). The three upper survival functions in this figure are quite similar, except the one for G_2 does not decrease as quickly for smaller values of t. The similarity is mostly due to failure mode R_9, which is included for each group and for which the most failure observations were available. The early difference is mostly due to failure mode R_6, which here affects units of groups G_1 and G_3, but not units of group G_2, and this failure mode caused several early failures. The main similarities and differences in these lower survival functions are due to the same effects. It is also interesting to consider when these three lower survival functions become equal to 0. For G_1, we have

$\underline{S}_{J_1^*}(t) = 0$ for $t \geq 7846$, because at this point the lower survival function for R_1 becomes 0, as for this failure mode only the data from group G_1 are used. For G_2, $\underline{S}_{J_2^*}(t) = 0$ for $t \geq 4329$, because at this point the lower survival function for R_2 becomes 0, as for this failure mode only the data from group G_2 are used. For G_3, we have $\underline{S}_{J_3^*}(t) = 0$ for $t \geq 13403$, because for this group the lower survival functions for R_5, R_6, R_9, R_{15} are used and these all only become 0 at this largest observation, as of course they were all observed for group G_3 so the data for this group are included.

It is also of interest to compare all NPI lower and upper survival functions in Figure 5, as this shows the effect of the different assumptions made, both with regard to the failure modes that will affect the future unit of interest and with regard to the data used for each failure mode, which is based on the assumption about which failure modes affected the units in the data groups. As mentioned, there is a considerable number of other combinations of assumptions, including to only use data from the single group of interest. The method presented in this paper can be used for all such inferences. Which assumptions are appropriate in a specific application necessarily requires the judgements of the topic experts. The approach can also be generalized, in a quite straightforward manner, to situations where the set of failure modes affecting the future unit differs from such sets that applied to units in the data, for example if a failure mode has been removed or if a new failure mode becomes relevant [8].

Acknowledgement Tahani Coolen-Maturi thanks the Institute of Hazard Risk and Resilience at Durham University for financial support.

References

[1] Augustin T. and Coolen F.P.A. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124** 251–272 (2004).

[2] Coolen F.P.A. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, **15** 21–47 (2006).

[3] Coolen F.P.A. Nonparametric predictive inference. In: M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, pp. 968–970 (2011).

[4] Coolen F.P.A., Troffaes, M.C. and Augustin T. Imprecise probability. In: M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, pp. 645–648 (2011).

[5] Coolen F.P.A., Coolen-Schrijner P. and Yan K.J. Nonparametric predictive inference in reliability. *Reliability Engineering & System Safety*, **78** 185–193 (2002).

[6] Coolen F.P.A. and Utkin L.V. Imprecise reliability. In: M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, pp. 649–650 (2011).

[7] Coolen F.P.A. and Yan K.J. Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, **126** 25–54 (2004).

[8] Coolen-Maturi T. and Coolen F.P.A. Unobserved, re-defined, unknown or removed failure modes in competing risks. *Journal of Risk and Reliability*, **225** 461–474 (2011).

[9] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. Nonparametric predictive multiple comparisons of lifetime data. *Communications in Statistics - Theory and Methods*, **41** 4164–4181 (2012).

[10] De Finetti B. *Theory of Probability*. Wiley, London (1974).

[11] Hill B.M. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63** 677–691 (1968).

[12] Kaplan E.L. and Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53** 457–481 (1958).

[13] Lawless J.F. *Statistical Models and Methods for Lifetime Data*. Wiley, Hoboken, N.J. (2003).

[14] Maturi T.A., Coolen-Schrijner P. and Coolen F.P.A. Nonparametric predictive inference for competing risks. *Journal of Risk and Reliability*, **224** 11–26 (2010).

[15] Walley P. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London (1991).

[16] Weichselberger K. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung: I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physica, Heidelberg (2001).

[17] Yan K.J. *Nonparametric Predictive Inference with Right-Censored Data*. PhD thesis, Durham University (2002). Available from www.npi-statistics.com.

Towards a Failsafe Flight Envelope Protection: The Recovery Shield

J.A. Stoop
Lund University, Sweden
Delft University of Technology, faculty of Aerospace Engineering,
Kluyverweg 1, 2629 HS Delft, the Netherlands

Abstract

Stall is an inherent unsafe flight state that has caused catastrophes in civil aviation. Solutions were gradually introduced, such as aerodynamic devices, pilot training, stick shakers and stall warnings. These mitigating, stand alone solutions seemed to have solved the problem to an acceptable level. However, stall as a phenomenon has recurred in a new form, due to changes in operating conditions, flight envelope protection, cockpit automation and crew competence qualifications. Stall accidents, as unintentional loss of pitch control, deal with the malfunctioning of primary flight control surfaces, automation mismanagement and hostile environmental influences during all weather operations.
As pitch control is the only primary flight control that has no design redundancy, a new control recovery approach is proposed, introducing new and uncorrupted aerodynamic forces by combining a technical recovery device with enhancing diagnostic abilities for the crew. In designing such a device, synchronizing time required and time available for flight control recovery is a critical design dimension This contribution elaborates on a control recovery device as an integrated system, consisting of a technical design device, computerized flight control and crew qualifications. A practical application of such a device focuses on stall recovery, as a specific category of flight control recovery.

1. Introduction

From the early days of aviation, stall has been an inherent hazard. Otto Lilienthal crashed and perished in 1896 as a result of stall. Wilbur Wright encountered stall for the first time in 1901, flying his second glider. Over the following decades, stall has remained as a fundamental hazard in flying fixed wing aircraft. Stall is a condition in which the flow over the main wing separates at high angles of attack, hindering the aircraft to gain lift from the wings. Stall depends only on angle of attack, not on airspeed. Because a correlation exists between loss of lift and minimal airspeed, a "stall speed" is usually used in practice. This is the speed below which the airplane cannot create enough lift at maximum angle of attack to sustain its weight in 1g flight. Airspeed is often used as an indirect indicator of approaching stall conditions. The stall speed will vary depending on the airplane's weight, altitude, and configuration. Fixed-wing aircraft have been equipped with devices to prevent or postpone a stall, to make it less more severe or to make recovery easier. Stall is an umbrella concept that covers various scenarios, aircraft configurations and has seen a wide variety of dedicated solutions. In practice, stall may occur under various conditions, configurations

and can be caused by various failure modes. Consequently, various stall scenarios exist.

Although stall is a prominent issue in loss of flight control, loss of flight control as a wider issue is related to more phenomena than pilot performance [1]. Firstly, hostile operating conditions may occur, dealing with icing, weather conditions, wake turbulence, thunderstorms or micro bursts. Non-revenue flight operations, flight training, pilot competences and automation mismanagement represent a second category of pilot related contributing factors to loss of control. Technical malfunction of flight control surfaces by internal or external threats pose a third category of factors, sometimes with catastrophic consequences. Recovery from loss of control situations depends on the competences of the crew under these specific conditions and on the nature and extend of the technical damage and loss of integrity of the airplanes' flight control functions. If the available recovery resources are insufficient and the time required for recovery lacks, the event becomes unsurvivable [2].

1.1 Loss of control scenarios

A fixed-wing aircraft during a stall may experience buffeting or a change in attitude. Most aircraft are designed to have a benign stall with characteristics that will warn the pilot. Because air no longer flows smoothly over the wings during a stall, aileron control of roll becomes less effective and may incline the aircraft to enter into a spin. The dangerous aspect of a stall is a limited recovery capability due to a lack of altitude. Such stalls may cause accidents at a low altitude. At high altitude, upper and lower speed limitations become critical as the speed range reduces. The upper limit is defined by structural integrity demands, while the lower speed limits depend on air density. As stall is reached, the aircraft will start to descend and the nose will pitch down. Recovery from this stalled state involves the pilot's decreasing the angle of attack and increasing the air speed, until smooth air-flow over the wing is restored. The maneuver is normally quite safe and if correctly handled leads to only a small loss in altitude. During training, a pilot is required to demonstrate competency to recognize, avoid, and recover from stalling the aircraft.
Finally, a loss of pitch control may occur due to other causes than aerodynamic stall, but may result in a stall. Such stall modes origin from technical failure of elevators, exceeding the allowable centre of gravity range due to shifting cargo or fuel imbalances, damage to elevators by external impact from space debris or bird strikes, loss of critical air data due to pitot port blockage by frost, foreign objects or insect intrusions. Despite all efforts to reduce stall and deep stall to acceptable levels of occurrence, such events still happen occasionally in commercial aviation. They raise concern about their emerging complexity, dynamics and impact on public perception on safety of aviation. Such events have been subjected to major accident investigations and serve as triggers for change throughout the aviation industry.

2. Towards loss of flight control prevention?

Recent major accidents indicate the potential for loss of flight control, at a high altitude as well as low altitude. A timely recognition is critical, but may be hampered by the ability of the crew to recognize and diagnose an event in a timely manner and respond accordingly, based on available time and resources. Transparency of the automated flight management system for the crew is a safety critical factor in the ability to diagnose and recover from an event. Under all flight circumstances a stable and controllable flight performance should be maintained.

2.1 Case histories

Air France AF447 crash
The BEA report on the AF447 accident demonstrates the complexity and dynamics of man-machine interfacing in which a continuous adaptation to a rapidly changing information display had to be taken into account. The eventual crash results from a succession of events in which [3]:

- A temporary obstruction of the pitot tubes, created inconsistencies in air speed measurements that caused autopilot disconnection and reconfiguration of normal flight control law mode towards alternate law mode
- Inappropriate pilot input destabilized the flight path
- The lack of linking between loss of indicated airspeed and appropriate crew procedures and the late identification of deviation from the flight path and insufficient correction applied by the PF
- The crew did not identify the approach to stall, their lack of immediate response and exit from the flight envelope and the crew failure to diagnose the stall situation and lack of inputs to enable a recovery.

In response to the obstruction of the pitot tubes by ice crystals, various monitoring systems triggered almost instantaneously. The crew is only informed of the consequences of the triggering by observing the disconnection of the automated pilot and the automated throttle and the shift to alternate law Electronic Centralized Aircraft Monitoring (ECAM). No failure message is provided that identifies the origin of these failures, in particular the rejection of the ADR's and of the speed measurements. No ECAM message enabled the crew to perform a rapid diagnosis of the situation, initiating the appropriate procedures. However, the crew is trained to read the ECAM as soon as the flight path is controlled, in order to analyze the situation and to organize a course of action to deal with the failures. Between the disconnection of the autopilot and the stall warning, numerous messages were displayed on the ECAM, but none helped the crew to identify the problem with the anomalous airspeed. Furthermore, the rapid change-over of the displayed information which was created by the flight computer in managing the priorities further complicated the crew's analysis and understanding of the situation. The reading of the ECAM by the pilots was time

consuming and used up mental resources to the detriment of handling the problem and monitoring the flight path.

Qantas Flight 32
The accident occurred at 10.01 hrs am by an uncontained failure of the port inboard (Number 2) engine, while en route from Singapore over Batam Island, Indonesia [4].
Debris from the exploding engine punctured part of the wing and damaged the fuel system causing leaks, disabled one hydraulic system and the anti-lock brakes and caused No.1 and No.4 engines to go into a 'degraded' mode, damaged landing flaps and the controls for the outer left No.1 engine.
The crew, after finding the plane controllable, decided to fly a racetrack holding pattern close to Singapore Changi Airport while assessing the status of the aircraft. It took 50 minutes to complete this initial assessment. The First Officer (FO) and Supervising Check Captain (SCC) then input the plane's status to the landing distance performance application (LDPA) for a landing 50 tonnes over maximum landing weight at Changi. Based on these inputs the LDPA could not calculate a landing distance. After discussion the crew elected to remove inputs related to a wet runway, in the knowledge that the runway was dry. The LDPA then returned the information that the landing was feasible with 100 metres of runway remaining. The flight then returned to Singapore Changi Airport, landing safely after the crew extended the landing gear by a gravity drop emergency extension system, at 11:45 hrs am Singapore time. As a result of the aircraft landing 35 knots faster than normal, four tyres were blown.
Upon landing, the crew were unable to shut down the No.1 engine, which had to be doused by emergency crews 3 hours after landing until flame out. The pilots considered whether to evacuate the plane immediately after landing as fuel was leaking from the left wing onto the brakes, which were extremely hot from maximum braking. The SCC pilot noted that in a situation where there is fuel, hot brakes and an engine could not be shut down, the safest place was on board the aircraft until such time as things changed. The cabin crew had an alert phase the whole time through ready to evacuate, open doors, inflate slides at any moment. As time went by, that danger abated and the crew was lucky enough to get everybody off very calmly and very methodically through one set of stairs. The plane was on battery power and had to contend with only one VHF radio to coordinate emergency procedure with the local fire crew.
There were no injuries reported among the 440 passengers and 29 crew on board the plane.

2.2 Emergency monitoring management

Effective use of checklists and ECAM message handling by flight crews during emergency and abnormal situations is extremely challenging. Using such interfaces is performed under extreme working conditions and very high mental work load relatively little guidance is available from human factor specialists, so during design and development aircraft designer have to fallback on historical

precedents, expert opinion and operator experience [5]. The linear nature of checklist and ECAM handling is very time consuming. The operators have to deal simultaneously with responding in detail to system malfunction indicators and maintaining oversight over the flight process. Preprocessing and selective representation of data by the message structure in a time constrained environment puts very high demands on the design of the man-machine interfacing. While handling loss of control situations, the levels of integration, automation and complexity place great cognitive demands on the flight crew. Such complex interfacing may create loss of situation awareness and mode confusion [6]. Structuring the interface along lines of human centered design is required to facilitate a supervisory role for the operator, because the dynamic behavior of the system directly influences the workload of the operator [7].
A conventional design creates the physical system first, then designs the control functions and identifies operator tasks. The interfaces as presented to the pilots are the only representation of the physical processes and flight control laws that are active and armed. Any malfunction of such interfacing or inappropriate responses and actions by the flight crew due to mode confusion becomes safety critical [6]. Due to the wide variety of operational situations and conditions, events may occur that deviate from the training and experience, hampering an accurate and timely understanding and response to the current and desired flight state. In most modern aircraft, Fly By Wire configurations flight computers are the only control path between pilot and flight control surfaces. The reliability and redundancy of such systems are safety critical. In the old days, where flight controls were handled manually, redundancy was provided by the skills of the pilots, gained by active flying training and experience. Today, more and more concern is expressed that basic flying skills have eroded by automation to the detriment of recovery capacity in emergency situations.

Several major events (AF447, QF32) indicate the thin line between successful skilled professional responses and a catastrophic outcome in a Fly By Wire environment. The classic notion of ‘human error’ as undesirable deviation from a normative concept of flight control is predominant among human factor specialists. Human error is commonly accepted among psychologists as the leading artifact in causing accidents. Human error should degrade the system from its optimal performance, creating mishap that could be prevented by safety management interventions. For psychologists, the rejection of the concept of ‘human error’ is difficult to rationalize with the perspective of the system designer employing a formal prediction methodology to help avoid actions that will degrade the system. *When considering human error, first of all pick your perspective then choose your label [8, pp 100].*
Consequently, automation would be the solution to human error, resulting in full automated flight. In this concept, there is no space for a critical reflection on the design of a supervisory role and discretionary responsibility of the pilot [7, 9]. Leaving aviation, navigation, communication to automated systems, pilots should restrict themselves to a managerial responsibility, balancing safety against efficiency and costs [8]. However, such concepts rely on almost flawless

automation and extreme low failure probabilities, irrespective of technological imperfections and harsh operating conditions. In practice, such an approach might not be the most appropriate perspective to analyze and understand complex and dynamic interactions between flight management systems and operators [10, 11]. It is a question whether it is possible to incorporate the know-how of operator experience into the design of safer systems [12]. Such a design could preserve craftsmanship and native resilience of such systems, relying on a high level of adaptability and professional expertise of the operators. These studies indicate potential adverse effects of classical safety interventions in terms of professional reluctance to accept further automation or through the emergence of new risks [12]. Constraining operator behavior in order to improve safety makes systems more rigid to the detriment of self-managed safety.
Such a role of the pilot as supervisor with oversight and control over the aircraft fits in well with the delegated responsibility of operators in a global network with distributed control over the primary production processes based on the principle of good airmanship [11].

In conclusion, three main sources for failure or success can be identified from these case descriptions:

- The available time window to deal with the situations was critical. The AF447 event took only 263 sec from the beginning to the very end, while the QF32 event took 4 hours and 45 minutes before the crew could declare the situation to be safe.
- Understanding the complexity and dynamics of the event consumed many resources. Diagnosing the event was to the detriment of the primary task to control the flight path in the AF447 event, while additional crew resources enabled to a high extend a successful handling of the QF 32 event.
- The availability of resources, redundancy and flexibility in responses determined the outcome of the events to a very high extend. Regaining oversight over the situation by strictly following procedures and check lists on a compliance based level would not have helped an understanding of the situation due to the damage to the Flight Management System and the structural damage to the aircraft.

3. Fly by Wire

In developing next generation aviation systems, new concepts are explored at the level of Blended Wing Bodies, composites, self-healing materials, flameless combustion, free flight, morphing wings and alike. This requires a different conceptual thinking than the classic man-machine-management interfacing. Reflection on common notions such as the usefulness of further automation is one of them. While fly by wire configurations were initially designed to reduce mental workload for pilots, modern developments have shown a much wider application of information technology and automation in the most modern generations of aircraft. How did FBW emerge and how did it evolve?

Automation was introduced in aircraft to perform two functions: eliminating pilot input as a source of error and automatically stabilizing aircraft under various flight conditions, tailored to various flight phases and system states. Fly By Wire interposed computers between the operator and the control actuators and control surfaces. Several control modes are available: normal, alternate, direct and mechanical. In older aircraft, control systems were mechanically and hydraulic redundant. Their characteristics limited the ability to compensate for aerodynamic conditions such as stability and structural properties while preventing unsafe flying states such as stall, spin or pilot induced oscillation. Redundancy was provided by pilot skills, gained by active flying training and experience. In order for a pilot to manually control stability, a nose downward rotating pitch momentum is required. While in older aircraft pilots were connected to rudder forces by haptic feedback, in new aircraft flight control laws serve two functions: inform the pilot to stay within limits of the flight envelope and take over the assessment of performance in case of failure. Which control laws are in force is aircraft state dependent, while mode selection depends on the flight phase: ground mode, flight mode, flare mode, each with specific protections regarding load factor limitations, attitude protection, high angle of attack protection, high speed warning and low energy warnings.

Stability of the aircraft is provided by gyroscopes for controlling pitch, roll and yaw changes. Since the pilot is not allowed to operate the aircraft out of the performance envelope, no natural stability is required. The pitch momentum coefficient can be deliberately low or negative to avoid a trim drag penalty on aerodynamic balancing of the aircraft. Such a negative coefficient however requires a rapid response to changes in pitch and attitude.

Since the Fly By Wire is the ONLY control path between pilot and flight control surfaces and computers do not fail by graceful degradation but instantaneously, redundancy is built into hardware and independent communication channels. For the regulatory and safety authority, a very low probability of failure was deemed acceptable, accompanied by vigorous testing and certification of components, followed by ultimately flight testing of the whole aircraft to get an Airworthiness Certificate. Installation of triple, quadruple equipment, separation of data buses, reboot and turnoff of faulty equipment combined with backup computers with reduced functionality takes over in case of total flight control failure.

A conventional backup was provided by override of the FBW equipment by the pilot. FBW has proven its value in military aircraft by reducing the workload of pilots, enabling a focus on their combat missions, keeping aerodynamic instable aircraft flying in a usable manner. In commercial applications, the handling characteristics were kept in within limits with respect to stall, spin, pilot induced oscillations, g-forces and structural loads. Computer protection was provided against undesirable handling in case of emergency situations such as ground proximity, loss of separation, wake vortex handling and collision avoidance maneuvers. Such protection against undesirable handling not only increases safety but also prevents over-engineering to protect against overstressing the structure, reducing the weight of the aircraft.

Gradually, FBW applications expanded to other flight control functions, based on the benefits they provided towards engine control, economy, maintenance, safety and performance. FBW enabled improvement of fuel economy by more accurate engine control on throttle setting, improving thrust balancing. Fuel trim optimization improved center of gravity control, eliminating trim rudder deflection and parasite aerodynamic drag. FBW reduced maintenance costs by eliminating analogue controlled actuator systems, simultaneously reducing weight of the aircraft. Intelligent compensation for damage and failure became possible by flight control law reconfigurations. Under normal law, FBW provides the ultimate flight control, while in emergency situations backup is foreseen by the computer shifting into alternate laws (Airbus) or override by the pilot (Boeing).
FBW technologies are vulnerable to failure because input sensors are single technology dependent: the altimeter is based on radio technology, while the pitot tubes depend on aerodynamic pressure. These dependencies are compensated by internal redundancy and a very low probability of failure, deemed acceptable by the certification authorities. This dependency however is not fail safe: loopholes remain and although rarely, have occurred under specific conditions. As such, the flight envelope protection is not (yet) fail safe, leaving room for improvement.
However, in accordance with NTSB findings, flight envelope protection falls short in case of deficiencies in pilot skills. In the Colgan Air crash (2009) fatigue and inability to respond to stall warnings contributed to the crash. In the flight AA587 crash (2001) excessive use of rudder controls in response to wake turbulence caused tail separation. Automation complacency remains an issue.

4. Towards innovative designs: a control recovery device

A wide variety of pragmatic and dedicated solutions to stall have been developed. They vary from aerodynamic –wing twist, stall fences, vortex generators and vortilons-, mechanic -such as stick pushers and shakers-, pilot oriented -such as such as horns and vocal messages- up to stall recovery devices -such as ballistic parachutes and tail rockets-. Although they have achieved a high level of sophistication in stall mitigation and recovery, a more fundamental approach to loss of flight control avoidance could be developed in order to deal with systemic deficiencies in loss of flight control avoidance. A timely deployment of new resources creates recovery potential and facilitates operating in a new and safe flight control state [13].
An innovative solution should comply with two principles:

- physical control over the aerodynamic forces that are exercised on aircraft by manipulating aircraft stability generating surfaces
- dynamic control over the behavior of the aircraft by operating of the flight controls beyond the level of procedural flight performance.

Such an innovative solution should encompass:

- introducing new aerodynamic forces instead of manipulating existing forces
- introduction of such aerodynamic forces in uncorrupted air flow
- generating high pitching moments by small forces combined with long arms

- introducing correcting forces only in case of emergency
- a timely and fail safe deployment in a 4D operating environment
- supported by a dedicated computerized flight mode for deployment in emergency handling.

In particular dealing with stall, an innovative design is suggested, based on these principles of resilience by dynamic vehicle control [14, 15]. Such a design as depicted in fig 1 is called a 'recovery shield device' and consists of the following features:

- on the fuselage of the aircraft, control surfaces are located at the nose and tail section in order to minimize the size of the surfaces, providing the largest momentum arm to the center of gravity
- these surfaces are only deployed in case of an emergent unstable flight to eliminate parasite aerodynamic drag in normal flight conditions
- these surfaces can be operated by either select nose or tail shields or combine a nose and tail mode of operation to provide a stable flight performance, depending on the stall scenario, flight phase, aircraft configuration and operating conditions
- a recovery shield control system is integrated in the flight management system, supported by dedicated computer software, depending on the level of sophistication of the aircraft control systems
- As a commercial application, it may serve as a safety asset in creating safety performance beyond legally required performance standards.

Figure 1. Recovery shields at the nose end of a T-tailed aircraft

Such devices and their control systems should benefit from deploying satellite system by developing avionics applications for ground speed, acceleration,

altitude, positioning and flight attitude identification to provide redundancy in technology over the pitot static data supply. In addition, a direct angle of attack indicator in the cockpit display is preferred to inform the pilot on the actual flight attitude of the aircraft and system state during loss of control.

Introducing such a device intends to serve flight safety by further development of the flight envelope protection by the introduction of a recovery shield. The recovery device is based on three notions:

- redundancy. The implementation of a recovery function for pitch control is necessary because of the loss of aerodynamic forces on the aircraft by disruption of the air flow across the wing and empennage. In addition, malfunctioning of the regular control surfaces may occur due to external or internal damage, failure of control actuators or as collateral damage due to other malfunctions such as structural collapse. Such a recovery function focuses on technical redundancy. Additional redundancy is provided by an overlap between technical redundancy and enhanced emergency handling capacity of the pilot in the recovery control mode of the flight management system
- resilience. The decoupling of a tight relation between the aerodynamic center and center of gravity range of the whole aircraft can create a more flexible range for the aerodynamic center by adding two small eccentric forces, deployed by two small extractable control surfaces. A further optimization of the center of gravity range is possible beyond the conventional cg range, facilitating a more economic and flexible use of the aircraft. This device does not replace the elevators, but reduces their size, reducing weight and parasite trim drag. Such resilience focuses on performance efficiency and eventually may lead to reconfiguration of the aircraft geometry as foreseen in the EU Framework program of smart wing development or even further into new concepts such as the Beechcraft Starship and application of canard wings
- responsive. There is a growing concern in the pilot community with respect to the reduction of flying and emergency handling skills under automated flight conditions and continuing degree of automation. Such a transfer from pilot controlled recovery action to aircraft controlled recovery devices seems the only option for commercial aircraft in the absence of the powerful thrust vectoring which exists in military aviation. In such a strategy, a human centered design in maintaining overall control over the situation seems preferable over a fully automated solution. The focus is on redistribution of the decision authority between aircraft and pilot and requires careful design of the man-machine interfacing. Such a transfer is to be accompanied by a simulator training program. By making the aircraft-pilot interface more responsive to degraded flight conditions and emergency conditions, the aircraft becomes less dependent of fluctuations and unforeseen situations in normal conditions. Such a responsiveness may reduce planning continuation errors and procedural flight performance.

5. Conclusions

Improving the recovery from loss of flight control, an additional recovery strategy should be provided if containment within the flight performance protective envelope is impossible. Expansion of the flight envelope protection is necessary, by introducing new and uncorrupted aerodynamic forces, redundancy in pitch control surfaces and expansion of flight management parameters. The rate of excursion from the protection envelope, available crew qualifications, the nature and extend of mechanical damage to control surfaces and required recovery time will limit the potential for successful application of such a recovery device.
Assessment of the flight control recovery device as a feasible and desirable innovation should be done in the early phases of its conception. Feedback from operationally highly experienced people such as pilots and accident investigators provide insights in the actual responses of the system under specific conditions that cannot be covered by an exhaustive proactive survey during design and development. A multi-actor assessment should identify strengths and weaknesses, opportunities and threats of the device, providing a safety impact assessment before the concept is released for operations.
Otherwise, the cure could be worse than the cause.

Acknowledgements
The author wants to express his acknowledgements to Frederick Mohrmann, Arthur Dijkstra and Hans Mulder of Delft University of Technology for their constructive discussions and valuable suggestions during the conception of the paper.

References

1. Veilette, P., *Investigation and preventing the loss of control accident.* ISASI Forum, Part I, July-September 2012, 5/9 Part II, October-December 2012, 19-24.
2. Stoop J.A., (2011). *Timeliness, an investigators challenge.* Investigation – A shared process, ISASI 2011, 12-15 Sept, Salt Lake City Utah, USA
3. BEA. *Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France Flight AF447 Rio de Janeiro – Paris.* Bureau d'Enquete et Analyses pour la securite de l'aviation civile. Paris, July 2012
4. ATSB. *In-flight engine failure - Qantas, Airbus A380, VH-OQA, overhead Batam Island, Indonesia, 4 November 2010* Australian transport Safety Bureau (2010)
5. Burian K. *Design guidance for emergency and abnormal checklists in aviation.* Proceedings of the Human factor and Ergonomics Society. 50th Annual meeting, San Francisco., October 2006.
5. De Kroes J.L. *Commercial plane or flight simulator, adjustable fuselage control surface, computer program product and method.* Patent P96519NL00. Deposited on 10 Jan 2012

6. Miller S.P., Barber S., Carlson T., Lempia D., Tribble A., *A methodology for improving mode awareness in flight guidance design.* Rockwell Collins Inc, Cedar Rapids, Iowa, 0-7803-7367-/02 IEEE (2002)
7. Meech J.F. *Addressing operator erros in supervisory systems*. British Aerospace Sowerby Research Centre, International Conference on Information-Decision-Action Systems in complex organizations, 6-8 April 1992.
8. Harris D. *Human performance on the flight deck*. Ashgate 2011
9. Martins E. and Soares M. *Automation under suspicion – case flight AF-447 Air France.* Work 41 (2012) 222-224 DOI: 10.3233/WOR-2012-0160-22 IOS Press (2012)
10. Stoop J. and Dekker S. *Are safety investigations pro-active?* Special Issue Safety science 50 (2012) 1422-1430. Future challenges of accident investigation , some insights from the 33rd ESReDA Seminar. Safety science 50 (2012) 1422-1430.
11. J.A. Stoop. *Time as a safety critical system integrator for stall recovery in aviation.* Human Factors and Ergonomics Society Europe Chapter, Human Factors: a view from an integrative perspective. October 10-12, 2012, Toulouse, France
12. Morel G., Amalberti R. and Chauvin C. *Articulating the differences between safety and resilience: the decision-making process of professional sea-fishing skippers.* Human factors, Feb 2008 Vol 50 issue 1 (2008)
13. J.A. Stoop and J.L. de Kroes. *Stall shield devices, an innovative approach to stall prevention?* 3th Air Transport and Operations Symposium, 18-20 June 2012, Delft University of Technology, The Netherlands
14. Hollnagel E. and Woods D.. *Resilience engineering: Concepts and Precepts.* Aldershot: Ashgate Publishing Ltd (2006)
15. De Kroes J.L. *Commercial plane or flight simulator, adjustable fuselage control surface, computer program product and method.* Patent P96519NL00. deposited on 28 Dec 2011.

Fig 1 is taken from *Flight Path 2050, Europe's Vision for Aviation*. Report of the High Level Group on Aviation Research. European Union 2011.

The Art and the Science of Whole Life Costing

Andy Kirwan and Julian Williams
Network Rail

Network Rail is one of the biggest asset management companies in the UK. In railway terms, we have the oldest system in the world and one of the busiest networks in Europe, with more train services than France, and more than Spain, Switzerland, The Netherlands, Portugal and Norway combined. We also have one of the safest rail networks, second only to Luxembourg in Europe, and one of the fastest growing, with a 50% increase in passenger journeys over the past decade and 30% more freight expected in the next five years.

The welcome increase in the demand for rail services presents a major challenge to asset managers – the people who plan and deliver work on the infrastructure. Additional trains increase the rate of asset degradation, restrict the time for access to the track to undertake preventive or restorative work, and exacerbate delays when failures occur. In parallel, there is a relentless drive for cost efficiencies, meaning the extra work needs to be done by fewer people.

To meet these challenges, we have had to rethink the way we prioritise work, to implement technological solutions that identify potential failures before they occur, and to devolve decisions to teams with a local understanding of the assets and close proximity to customers. To support this shift, we have introduced a whole life costing framework that puts customer service at the centre of decision making and provides consistency across asset disciplines and between business functions.

In this presentation, we will describe the approach taken, explain how it has been practically implemented, and show how the results have informed our investment plans for the next five years. Emphasis will be given to the models we have developed, the influence of uncertainties on decision making, and the compromises that are necessary to integrate ‘top down’ forecasts with ‘bottom up’ real world plans.

Localising Risk Estimates from the RSSB SRM

Chris Harrison
RSSB, Block 2 Angel Square, 1 Torrens Street,
London EC1V 1NY, UK

Abstract

The Rail Safety and Standards Board (RSSB) Safety Risk Model (SRM) provides a structured representation of the causes and consequences of potential accidents arising from railway operations and maintenance on the UK railway network. It consists of a series of fault tree and event tree models representing 120 hazardous events that collectively define an overall level of risk on this system.

Recently there have been several RSSB led projects looking at how to take the national results from these models and apply them at a local level. This involves breaking the network up into smaller regions and using the localised data to estimate the risk. There are two main ways of doing this: either scaling the national profile using appropriate normalisation data or by taking the calculations used at the national level and applying them to the localised data.

The paper will look at the tools RSSB has developed over the years to carry out these assessments. The scaling of the SRM figures from a national profile to a localised profile is done via the SRM Risk Profile Tool. The application of the SRM calculations to localised data to calculate a localised profile can be done via a tool called the Network Modelling Framework (NMF) Safety Module. The purpose of this paper is to discuss how these tools work, their limitations, some of the challenges that have been encountered and current research aimed at developing a scalable ‘GeoSRM’.

1. Introduction

1.1 RSSB

Rail Safety and Standards Board (RSSB) was established on 1 April 2003, implementing one of the core sets of recommendations from the second part of Lord Cullen's public inquiry into the Ladbroke Grove train accident. The company's prime objective is to lead and facilitate the railway industry's work to achieve continuous improvement in the health and safety performance of the railways in Great Britain, and thus to facilitate the reduction of risk to passengers, employees and the affected public.

1.2 Aims of the Paper

The aims of this paper are to:

- Give an overview of the RSSB Safety Risk Model (SRM).

- Explain how localised risk profiles are currently generated using the Risk Profile Tool (RPT) and the Network Modelling Framework Safety Module (NMF SM) and discuss their limitations.
- Present an overview of work that is currently under development to create a new localised tool based on the SRM.

2. Outline of the RSSB Safety Risk Model

2.1 Purpose of the Model

RSSB has a responsibility to lead and develop long-term safety strategy and policy for the Great Britain railway network (GBRN). The SRM has been developed by RSSB to provide a structured representation of the causes and consequences of potential accidents arising from railway operations and maintenance on the mainline rail network.

2.2 Overview

The SRM consists of a series of fault tree and event tree models representing 120 hazardous events that collectively define an overall level of risk on the GBRN [Refs 1 & 2]. A hazardous event (HE) is defined as an event that can lead directly to death or injury. Each HE consists of a fault tree that models all of the precursors that can lead to, or contribute to, the HE. The top event of the fault tree is the initiating event of an event tree that models the consequences that can result from the HE.

The SRM relates to the average network risk on the GBRN covering all running lines, rolling stock types, locations and stations currently in use. The SRM has been designed to take full account of both the high-frequency low-consequence type events (events occurring routinely for which there is a significant quantity of recorded data), and the low-frequency high-consequence events (events occurring rarely for which there is little recorded data). The results for each HE are presented in terms of the frequency of occurrence (number of events/year) and the risk (number of injuries per year).

3. Current SRM Localisation Tools

The current methods used to localise the SRM risk estimates involves breaking the GB rail network up into smaller regions and using the localised data for these regions to estimate the risk. There are two main ways of doing this:

- Scaling the national profile using appropriate normalisation data
- Taking the calculations used at the national level and applying them to the localised data.

The first of these is done via the SRM Risk Profile Tool (RPT) while the second of these is done via a tool called the Network Modelling Framework (NMF) Safety Module [Ref 3]. The rest of this section will look at these two tools.

3.1 Risk Profile Tool (RPT)

The SRM RPT enables a user to produce a localised risk profile of their railway operation based on their usage and injury data. The main way in which this data is used is to scale the frequency of event for each precursor that makes up a hazardous event. Each of the SRM hazardous events has a *normaliser*, which for the national SRM represents one of the key dimensions of the railway (eg total passenger journeys, passenger train km). These are used on a proportioning basis, ie if company A carries out 5% of the passenger train miles then they will be responsible for 5% of the events that are normalised by passenger train miles. This assumes that the local operation has, on average, similar characteristics to the operation of the national mainline railway, which is quite a gross assumption given the network's heterogeneity.

However, the users can take this into account where necessary and adjust their risk profile to more closely model their own operation. This is done through the use of their injury and accident data to alter their frequency and consequence estimates. The tool allows these modified estimates to be used instead of the scaled down national averages, so as users can better reflect their own operation.

3.2 Network Modelling Framework (NMF) Safety Module

The NMF is a mathematical model that has been created by the Department for Transport (DfT) to analyse specific investment scenarios including a prediction of what effect these will have on safety for the GB railway industry.

The NMF model contains different modules for analysing different types of data. These include: a Safety Module (SM); Demand Module; Performance Module; and an Infrastructure Cost Module. RSSB was originally involved in the development of the SM for the first version of the NMF used to support railway investment decisions for Control Period 4 (2009 – 2014). The NMF was used to analyse a number of investment scenarios and the results from the SM assisted in determining the target risk reductions that could be expected from the investment. The resultant High Level Output Specification (HLOS) safety targets require a reduction in the normalised risk to passengers and workforce of 3% over Control Period 4.

The SM is designed to estimate how the safety risk profile will change on each Strategic Route Section (SRS) in future years given different investment scenarios. There are 285 SRSs modelled in total. These use the modelling structure and data from the SRM, supplemented with SRS specific data from the NMF input files to calculate a customised risk profile for each SRS. This is done by taking the calculations used at the national level and applying them to the localised data.

3.3 RPT and SM localised risk modelling limitations

One of the key limitations of the RPT assessments is that they are completed independently. That is to say, each user is responsible for completing their own assessment of their risk using their judgement where necessary to alter their profile. For this reason, it is no surprise that if the total of these assessments is calculated, it will be different to the total risk outputted by the SRM. The same is true of the SM, in that the sum of all of the SRS does not necessarily equate to the total overall SRM risk.

The other limiting factor of the RPT is that it is has been set up to proportion risk based on event frequencies and exposure metrics (via the normalisation data). It is difficult to use the tool to modify the average consequences per hazardous event, as the modelling structure to do this is not present in the tool. Where it is done it is mostly based on expert judgment and is another reason why if the outputs from all the RPTs were combined they would not match the national risk profile.

One of the key features of the SM is that it uses the same calculations as the national SRM, but uses local data rather than the national SRM averages to calculate the model parameters. The main problem with this sort of approach is one of consistency between the local parameters and the national SRM average parameters. It is desirable that equation (1) for all the parameters of the two models is satisfied.

$$X_{SRM} \cong \sum_{SRS} w.X_{SRS} \qquad (1)$$

For given parameter (X_{SRM}), the comparative value calculated via an appropriately weighted (w) combination of each of the SRS values of the parameter (X_{SRS}) should ideally be congruent, or if not then match within a given tolerance. In general for the SM in its current form, this is not necessarily always true and in some cases can be significantly different. The main reason for this is the non-linearity of some of the SRM equations along with the correlated structure the parameters are built in. This makes it very difficult to satisfy equality (1) for all parameters simultaneously, as calibrating one parameter can offset the calibration on another.

Another significant issue is that the national SRM parameters are calibrated against historical data and set so that nationally the numbers make sense. However, if these same equations and methods are applied locally, then the assumptions made at the national level may begin to break down. This problem lies with the interaction and correlation of many of the factors with one another. The main issue is how to successfully calibrate each of the 285 SRSs individually, whilst still retaining an overall average of them that is consistent with the national average.

The other key issue with localised risk modelling for both the RPT and the SM concerns the availability of data. When collated nationally there may be ample incident and accident data to confidently estimate the necessary

parameters of the model. However, as this data is sub-divided into ever smaller parts to represent routes or a particular railway operation, the problem of sparse data or even no data can become an issue.

There are various ways round this problem, either by utilising expert judgement to make estimates where there is little or no data, or by employing statistical techniques to apportion the known data accordingly. The main technique used in the SRM and the NMF SM is empirical Bayes estimation and Refs [4, 5, 6 & 7] give more detail about how this has been applied. The technique essentially provides a framework for pooling similar incident data together (including incidents that have not been observed) and determining a suitable weighted occurrence rate for each incident that is based on the average rate of the pool and the actual observed rate from the incident data.

4. Future Localised Risk Modelling using the SRM

The preceding sections have discussed how localised risk modelling is carried out using the SRM and the tools that have been developed to determine these profiles. This section will present RSSB's plans for the future and outline our vision and the tool we are developing to achieve this goal.

4.1 GeoSRM

RSSB is currently developing a GeoSRM, the aim of which is to provide geo-referenced risk estimates for safety hazards across the GB rail network. It will be based on the SRM and it is intended that it will show how risk is distributed across the network. To do this it will need to take account of many local factors, such as:

- Infrastructure assets
    - Track
    - Points
    - Signals
    - Stations
    - Bridges
    - Level crossings
    - Tunnels
- Geographical features
    - Embankments
    - Cuttings
- Train services
    - Train routing
    - Type of train
    - Train loading

The GeoSRM will be available to industry stakeholders (much like the current RPT) and it is intended to be made accessible via a web interface. The intention is for to allow users to submit queries on the risk level and see the resulting risk estimates displayed graphically over a map of the rail network. This will enable users to see how risk varies along a route and how variations in risk are associated with the local factors at a given location, route or

whatever level the estimates are required at. Figure 1 shows a diagram of a potential set-up for the GeoSRM and its various features.

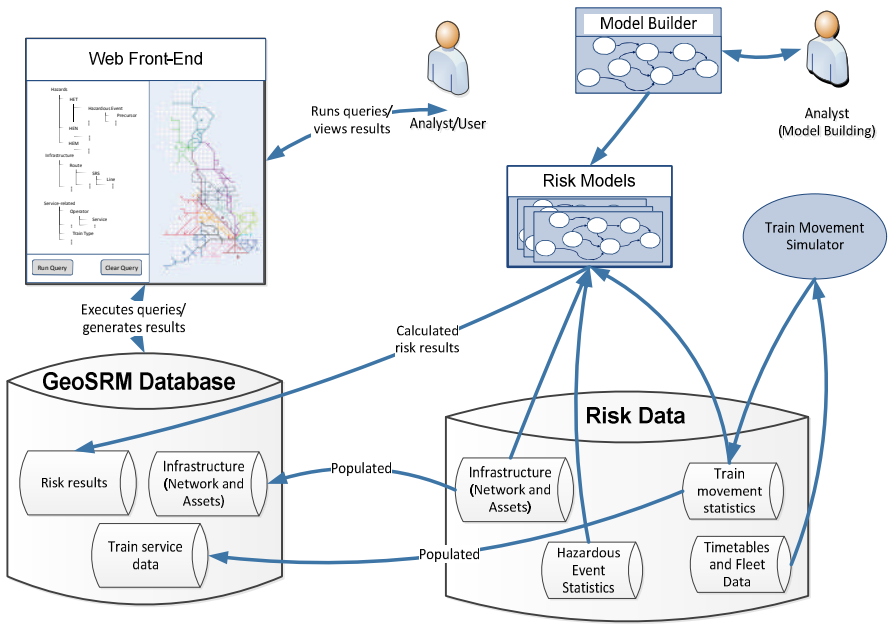


Figure 1. Schematic of the potential GeoSRM set-up

4.2 Issues to address

The GeoSRM is in many ways similar to the existing SRM localisation tools (RPT and NMF SM) and will have to overcome many of the problems highlighted in Section 3.3. The main issue will be how to confidently model risk at different levels of granularity, whilst still retaining consistency between different combinations of the outputs and the overall total.

Figure 3 gives an overview of some of the characteristics of the modelling at different levels. The arrows underneath represent some of the features of the modelling and the arrow indicates the direction in which the feature is increasing as you move through the different modelling levels.

Along the top, moving left to right, the modelling level becomes increasingly localised: moving from the national network, to a strategic route to a section of track between two stations. This can result in an increase in the amount of data required to carry out the analysis, both in terms of inputs and outputs, as you become more and more localised.

The main issue with this is that the amount of incident data on which to base the risk estimates becomes less and less as the level of localisation increases and this will lead to increasing statistical uncertainty. The main benefit, however, is that by further localising the risk estimates you can reduce the

modelling uncertainty and take into account more and more factors that affect the risk at the appropriate level at which they should be analysed[1].

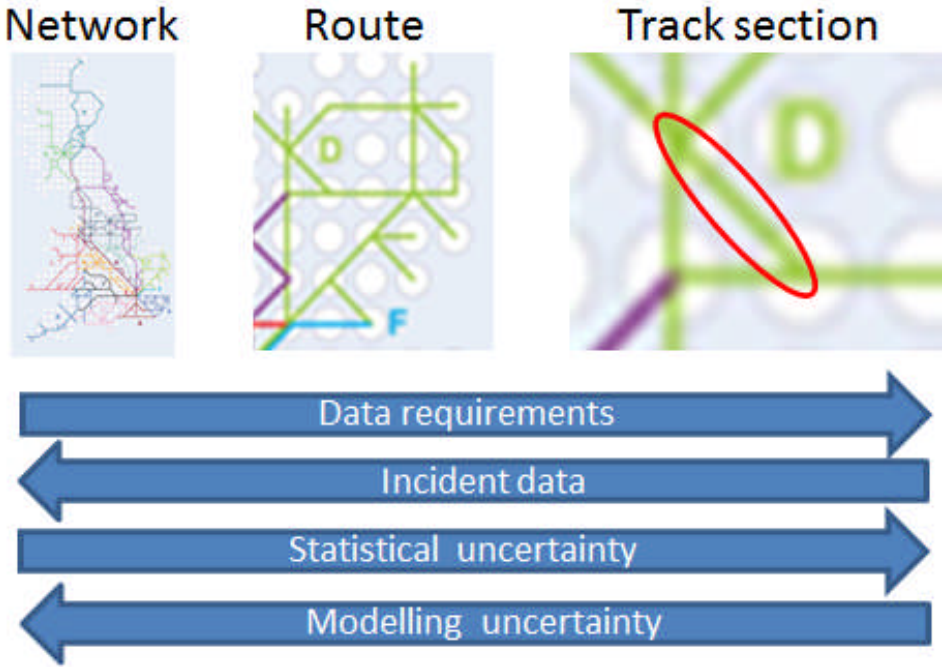


Figure 2. Characteristics of modelling at different levels

4.3 The solution?

In the course of the development of the GeoSRM, RSSB will have to contend with all of these features of the modelling and construct a framework that will enable consistent, localised risk estimates to be calculated based on the underlying characteristics of the level under investigation. The solution to this problem is likely to be a bottom-up approach to the risk estimation, which is very different to the current top-down approach taken. The degree to which this is undertaken is uncertain, however, it is likely to mean that the SRM parameters are built up from local profiles rather than the current way of working out national profiles and adjusting them down.

However, that does not mean that the top-down approach is no longer necessary. As Figure 2 illustrates, the more localised the risk modelling becomes, the less incident data and the more uncertainty there will be in the risk estimates. Therefore, initially at least, there will be the requirement to calibrate the bottom-up approach by consideration and comparison with the top-down approach. Eventually the aim would be to have sufficient confidence in the bottom-up approach to forego the top-down approach and use the GeoSRM as a tool to assess and analyse the GB national rail risk profile.

[1] Take for example the SRM estimates for the probability of a train colliding with a derailed train blocking an adjacent line. These are currently based on calculations of train passing frequencies carried out on the national rail timetable. It would be more accurate to base these figures on more localised sections of track and the train service on them and then recombine these to give a national figure.

4.4 GeoSRM Implementation Programme

Work is currently under way to complete a pilot study of the GeoSRM. The aim of this work is to create a working model along with a prototype architecture of the tool and apply it to part of the GB rail network. This will involve creating a database and a front end for the tool that will allow risk to be analysed and visualised as outlined in the preceding sections.

The pilot study will only consider three hazardous events, namely:

- Slips, trips and falls at stations
- Derailments
- Suicides

The reason for selecting these is that they provide a broad cross-section of hazardous events that are geographically dispersed and which will allow diverse datasets to be incorporated into the analysis. The pilot study is currently underway and initial results are expected to be available in late summer 2013. After that it is anticipated that work can progress to development of a GeoSRM representing the entire GB network for all hazardous events currently modelled by the SRM.

5. Summary

This paper has outlined the current localisation tools that RSSB uses to assess rail risk. The limitations of these tools have been discussed and a new tool, the GeoSRM, has been introduced. Some of the potential problems that may be encountered in the development of this tool have been highlighted and RSSB's current thinking on how to address these has been summarised.

The main issue to contend with is one of congruence and how the models are constructed and the analysis carried out to ensure that the localised assessments make sense in the context of the overall risk profile.

References

1. Bearfield G. et al, Safety Risk Model Risk Profile Bulletin, version 7, RSSB, (2011).
   www.safetyriskmodel.co.uk

2. Dennis C. and Somaiya K., Development and Use of the UK Railway Network's Safety Risk Model, Probabilistic Safety Assessment and Management: PSAM 7 - ESREL '04, (2004).

3. RSSB, T956: Further development of the Department for Transport Network Modelling Framework safety module, RSSB, (2013).

4. Bedford T., Quigley J. and Walls L., T172, A statistical review of the RSSB Safety Risk Model: WP2 Report, Strathclyde University, RSSB, (2004).
   http://www.rssb.co.uk/SiteCollectionDocuments/pdf/reports/Research/T127_rpt2_final.pdf

5. Harrison C., The use of Empirical Bayes Methods in the Rail Safety & Standards Board (RSSB) Safety Risk Model, 27th ESReDA, Glasgow, (2004).

6. Quigley J., Bedford T. and Walls L., Estimating rate of occurrence of rare events with empirical Bayes: A railway application, in *Reliability Engineering & System Safety*, Volume 92, Issue 5, pp619-627, (2007).

7. Griffin D. and Holloway A., Route based risk estimation across the GB rail network using empirical Bayes methods, ESREL 2012, Stockholm, (2012).

Use of a Generic Hazard List to support the development of Re-usable Safety Arguments in the Rail Industry

George Bearfield, Reuben McDonald
Safety Knowledge and Planning Department, RSSB, London, UK
Angel Square, One Torrens Street, London, UK

Abstract

The paper describes an ongoing research project to produce tools and guidance to support the rail industry in undertaking engineering safety management activity in accordance and clear compliance with the requirements of the European railway regulation for the 'Common Safety Method on Risk Evaluation and Assessment' (CSM on RE&A) [6].
A key aspect of this work is the development of generic hazard lists to support:
i Demonstrable application of the CSM on RE&A.
ii. Use of the process to support the consistent management and transfer of hazards between different parties involved in a significant change project relating to rolling stock.

The research project has developed a draft generic hazard list for rolling stock hazards. This followed a review of several hazard lists produced by manufacturers and operators. This hazard list is being used as a consistent basis for risk assessment and management processes that are performed under the framework.

RSSB has identified the 'Codes of Practice' (eg in Railway Group Standards, and European standards) that form part of the 'System Definition' or that provide mitigation of the risk of certain causes of the generic hazards in accordance with the framework. Using a generic system definition various mappings are being produced from the 'System Definition', to causes of the hazards, to the safety requirements. The intent is for these mappings to support a template approach to application of the CSM with fragments of the argument being able to be re-used and adapted over time as successive projects prove their validity. This will ensure the safety work is both cost effective and suitable controls and their importance are appropriately identified.

1. Introduction

This paper describes work undertaken under RSSB research project T955, on behalf of the GB railway industry. The objective of project T955 is to investigate the extent to which a common set of definitions, processes and tools for hazard identification analysis and safety risk management would benefit the rail industry, and to develop such definitions, processes and tools where a benefit is identified. Phase one of the project included interviews with stakeholders across the rail industry. Phase two of the project seeks to develop guidance and tools so they may be used by the rail industry.

2. The CSM on RE&A

The Common Safety Method on Risk Evaluation and Assessment [6] (hereafter known as the CSM on RE&A) is a European regulation that is part of a wide-ranging programme of work by the European Commission to bring about a more open, competitive rail market while ensuring that safety levels are maintained, and, if reasonably practicable, improved. The CSM proposes a framework for assessing and accepting 'significant changes' on the railway based on the analysis and evaluation of hazards. The framework itself is broadly consistent with the principles of system safety that are common across a range of industries, and embedded in the Railway Industry through standards like EN50126-9 [2-5] and, prior to its withdrawal, the 'Yellow Book' [6]. A key difference arises in the principles for accepting risk. In each European member state risk acceptance was previously solely dependent on national legislation which differed from member state to member state. The CSM brings about greater commonality in the principles for accepting risk by proposing the use one of three risk acceptance principles for demonstrating that the risk from hazards has been sufficiently addressed:

- Application of codes of practice
- Comparison with similar systems (reference systems)
- Explicit risk estimation

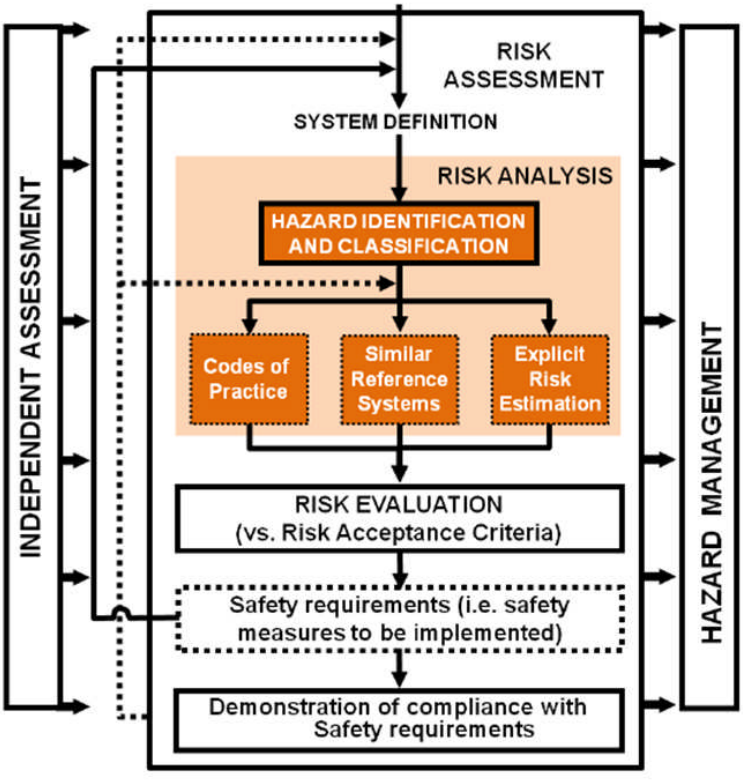


Figure 1. Diagram summarising the CSM on RE&A

This change in regulation provided the potential for reviewing the GB approach to optimise the efficiency of safety approvals within this new framework.

RSSB has initiated a research project on behalf of, and with the support of, the GB railway industry. The next two sections of this paper describe the two different sequential phases of that project, their findings and outputs.

3. Phase one: the review of current practice

The first phase of this work [8] undertook a survey of how the various elements of the CSM on RE&A were being applied in practice by the various actors in the GB railway industry. A questionnaire was designed to elicit a view of the current state of practice in safety engineering and analysis and the potential impact of the CSM on RE&A on projects in the GB railway industry. Responses were invited from safety practitioners in Network Rail, London Underground, Crossrail and the supply and consulting sectors of the industry. A range of issues were found, but in this paper the focus is on the findings relating to the potential for a more standardised approach to hazard management (eg hazard identification, analysis, transfer and closure).

The two key findings in this regard were:

- Many respondents noted that the term system is an ambiguous term, for example:
  'People use 'system' without defining what they mean. When dividing the system into subsystems it can be difficult to determine to which subsystem things belong, and also to classify things which do not clearly belong to a particular subsystem'.
- Significant differences were found in the classification of things as hazards, causal factors, barriers and accidents. This relates to the differing understanding of a system boundary as well as the differing approaches to classification of 'causes' as hazards. It is clear that there is currently no consensus over the distinction between different types of hazards and causes in particular for the operational railway as a whole (as opposed to the systems within it). This potentially creates issues with the size of hazard logs, the repeatability of analysis work from one project to the next, and difficulties with communication of safety related information between individuals and across organisational boundaries throughout the system lifecycle.

On the basis of the finding above, in the second phase of work the development of a generic hazard list has been undertaken to improve the identified inaccuracies.

3. Phase two: the evolving framework

There are many aspects to the evolving framework of tools and guidance being developed by the project. In this section we describe key elements of the framework with regard to hazard management.

2.1 The concept of using a generic hazard list

Hazards are traditionally specified as those at the boundary of responsibility of the railway company (transport undertaking or infrastructure manager), however within the contact of the CSM on RA&E, ultimately the hazards that must be controlled are those at the boundary of the railway as a system. The generic hazards developed as part of T955 are compliant with the definition in the CSM on RA&E regulation [6] of a condition that could lead to an accident and are thesefore defined at the railway system level.

Using a common generic hazard list supports effective communication between the different parties in the supply chain. The hazards and controls are identified by the proposer of the change (normally the transport undertaking / infrastructure manager) and then hazards or safety requirements (linked to specific hazards) cascaded though to the suppliers. In return the evidence is supplied back up the chain and is linked directly to the appropriate hazard in a way which is meaningful and supports a clear safety demonstration.

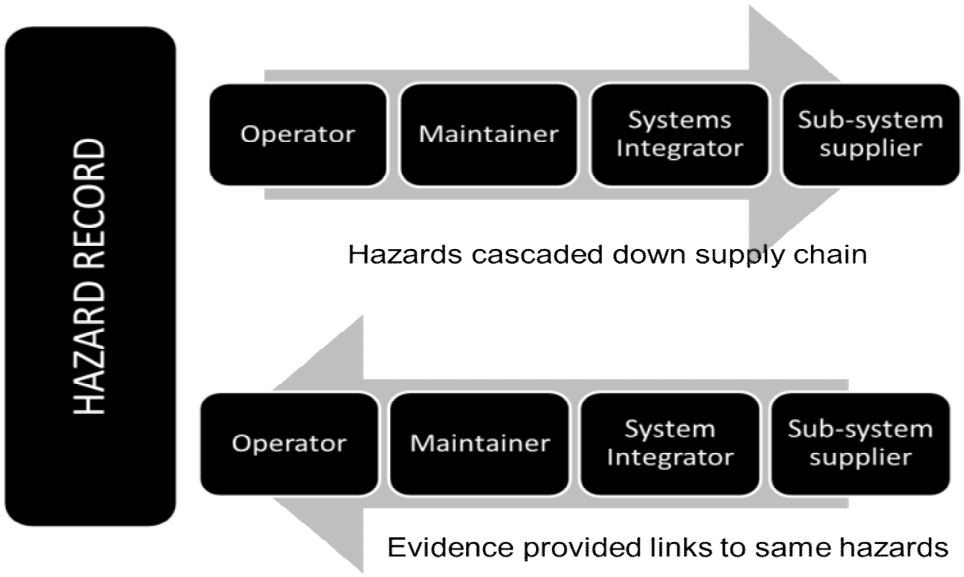


Figure 2. Cascading the generic hazard list through the supply chain, example shown is for a Transport Undertaking such as a train operator contracting for the supply of new equipment.

2.2 System Definition and Hazard Identification

Within the context of the CSM on RE&A the significant change being considered is of the ‘railway system’. Note that this encompasses the entirety of the technical functions and procedures required to deliver a working railway in addition to its operational context and physical environment. The safety directive states that the system definition should address at least the following issues:

- System objective, eg intended purpose.

- System functions and elements, where relevant (including eg human, technical and operational elements).
- System boundary including other interacting systems.
- Physical (ie interacting systems) and functional (ie functional input and output) interfaces.
- System environment (eg energy and thermal flow, shocks, vibrations, electromagnetic interference, operational use).
- Existing safety measures and, after iterations, definition of the safety requirements identified by the risk assessment process.
- Assumptions which shall determine the limits for the risk assessment.

The system definition is used to support the hazard identification exercise and in effect all of these various elements need to be considered to the extent that they influence the occurrence of hazards and the risk that these hazards present.

In we consider the risks of changing the method of train dispatch, we would need to consider the scenario of a train being dispatched from a station we would need to define:

- Functions such as ‘There is a speed interlock at 3 kilometres per hour which prevents doors being opened above this speed’.
- Operational procedures such as ‘After the driver has closed the doors, they must check that the door interlock light is lit’.
- Operating context ‘the dwell time of the train is typically 5 minutes at the peak service’.
- Environmental: the station is close to a town centre and at certain times of the week many of the passengers are likely to be intoxicated.

2.3 Hazard Identification and use of the Generic Hazard List

The use of the generic hazard list for hazard identification depends if a similar change has been through the process before and a template argument exists. The twin uses are described below in Figure 3.

If a similar change has not been through the process before then the hazard list can be used to support hazard identification for the change. In the approach being developed the hazard identification will look at a certain ‘phase of mission’ eg the processes and functions for despatching a train at a station. Therefore it is sensible that the system definition is presented in this way to enable it to fully support the hazard identification exercise.

A systematic analysis of the system definition can be performed using appropriate experts in a workshop environment facilitated by use of the generic hazard list. For example, using a chronological sequence of each technical function and procedure and categorising hazards using the list; the list can also be used as part of a completeness check later on in the workshop to ensure all known hazards have been considered.

If a template argument is available, the generic hazard list can be used as the basis of the new assessment. A detailed gap analysis would need to be undertaken to compare the system definition for the template argument and new assessment and any additional causes or hazards identified.

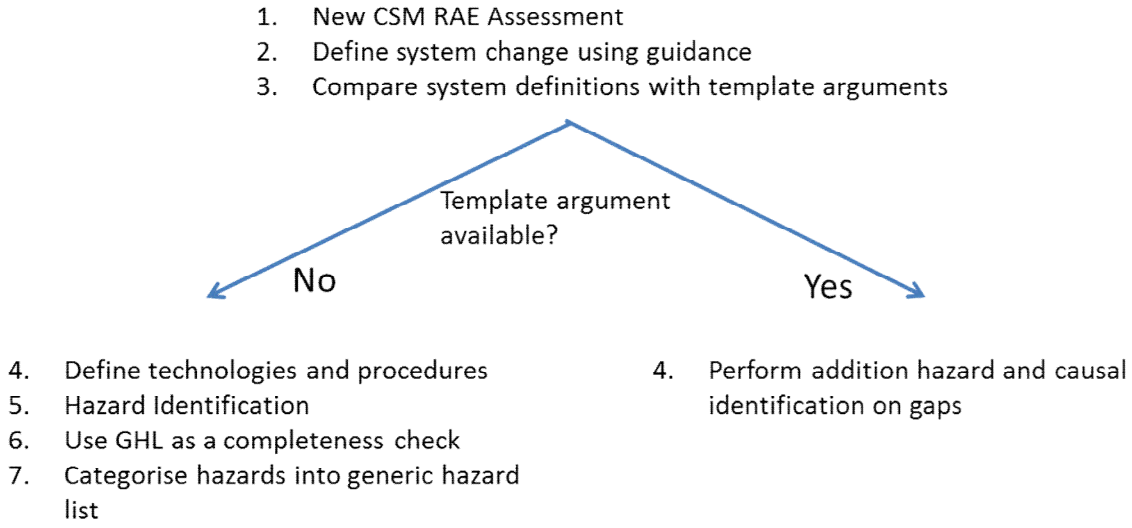


Figure 3: Process for the use of the Generic hazard list for hazard identification, depending if a similar change has been analysed and published as a template argument

2.4 Joining up to Codes of Practice and Other Safety Evidence

The hazard management process is undertaken with the purpose of ensuring that the risk from all hazards identified has been managed to an acceptable level. The process being developed supports the clear demonstration of this and communication and transfer of hazards between various parties involved.

RSSB is mapping the 'Codes of Practice' (eg in Railway Group Standards, and European standards) that form part of the 'System Definition' or that provide mitigation of the risk of certain causes of the generic hazards in accordance with the framework.

Using a generic system definition various mappings are being produced from the 'System Definition', to causes of the hazards, to the safety requirements. The intent is for these mappings to support a template approach to application of the CSM with fragments of the argument being able to be re-used and adapted over time as successive projects prove their validity.

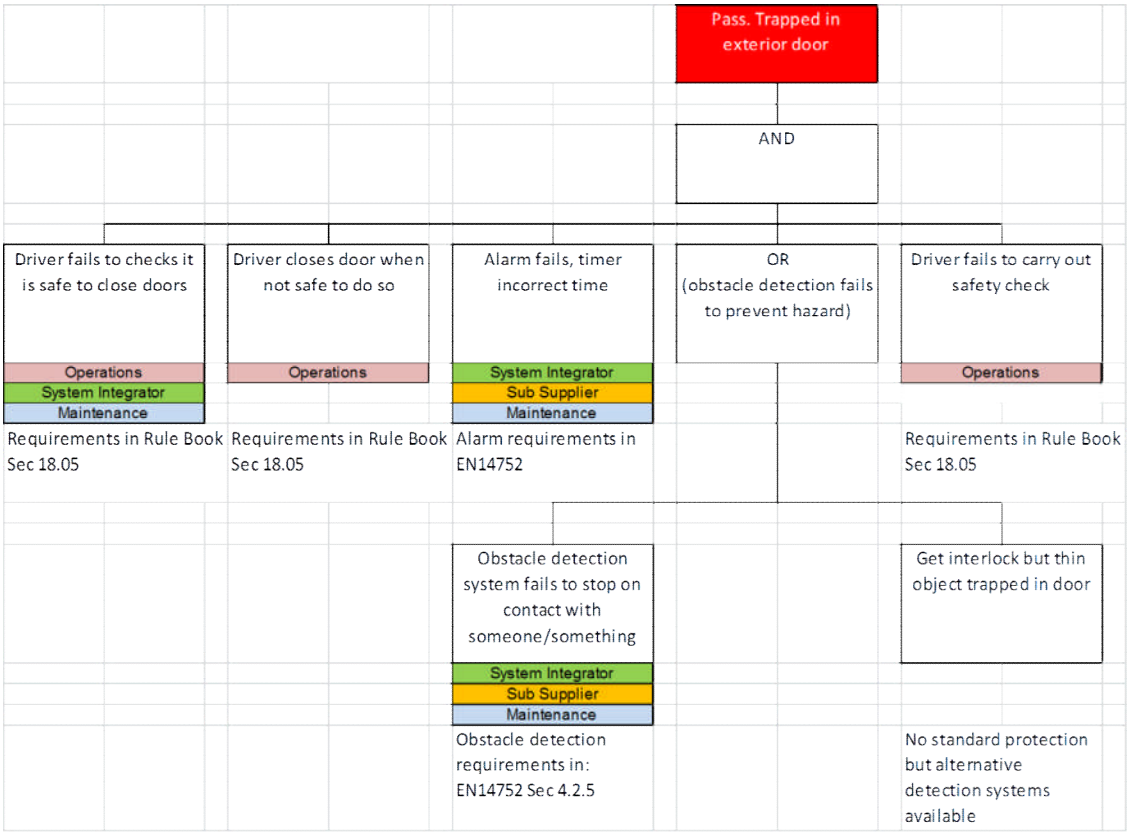


Figure 4. Generic hazard 'passenger trapped in exterior door' with associated causes and controls

Note, this is not just about codes of practice – the mapping should help link to any type of evidence eg native acceptance evidence, particular ALARP assessments etc. However, the Codes of Practice must be applied so an understanding of the extent to which they mitigate the hazard risk is needed.

The example given in Figure 4 describes the key controls associated with train dispatch and the hazard of a passenger trapped in an exterior door. These controls cover technical specifications, eg Euro-norms called by the appropriate TSI and operational controls such as RSSB rule book measures, In addition the actors involved in managing those causes are recorded.

2.5 Re-Use of Safety Evidence

It is intended that a number of assessments are performed for different common system definitions in real-world applications; these would be made anonymous and published by the RSSB.

The published assessments would provide a great advantage to future projects that are making similar changes to the railway system. They would provide a templated argument for a given system definition.

3. Future work

Initial work has focussed on development of the generic hazard list with a scope of hazards that affect rolling stock. Causes of those hazards and an initial set of controls in the form of codes of practice have been identified.

Work planned for 2013 is two-fold. Two initial case studies have been identified that will be used to test the generic hazard list process and to supply a causal breakdown and set of controls.

One of these is a technological change involving the introduction of ERTMS into a specific ~~scenario,~~scenario; the other is an operational change involving the change in role of a train driver. It is intended that these case studies will result in template safety arguments being available for future users to apply.

In addition to these case studies, a number of guidance documents are being produced to support CSM on RA&E assessments and in particular the use of the generic hazard list. These are summarised below in Figure 5.

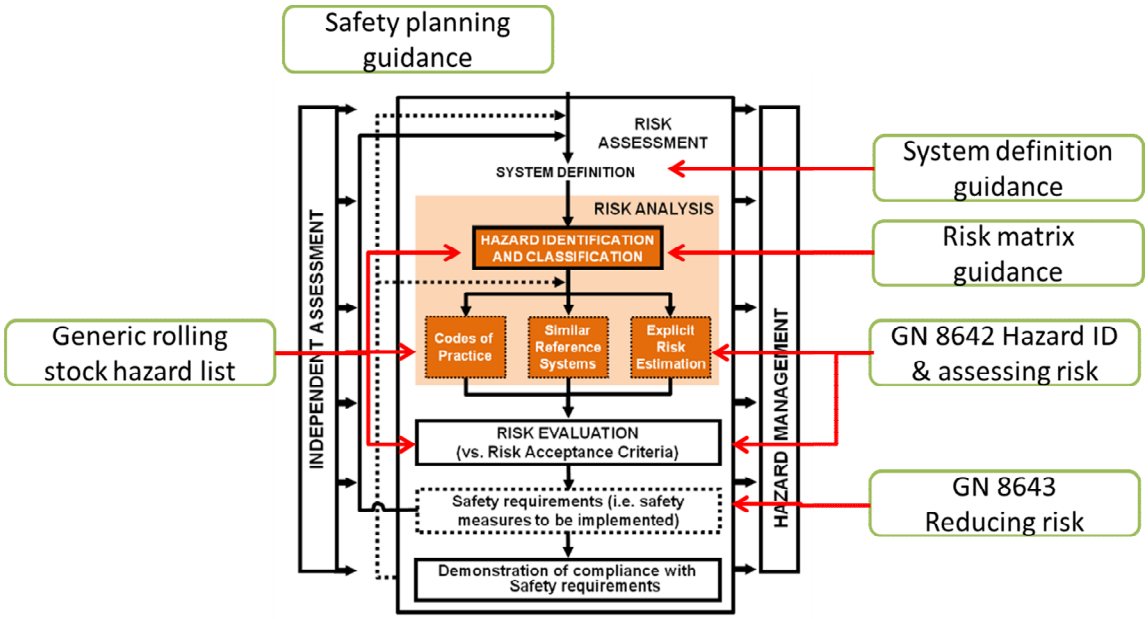


Figure 5. Guidance documents existing and planned as part of T955 project on the application of the CSM on RAE. GN 8642 and 8643 have already been published, other documents are planned for publication in 2013.

4. Related work

A generic hazard list for the functional safety of railway vehicles forms part of a system developed by the German Federal Railway Authority (EBA) called SIRF [9]. This differs in scope as it is primarily concerned with the functional safety, rather than considering all operational aspects. The draft 50126 standard [10] includes a railway wide example hazard list for use as the basis of hazard identification and consistent naming. This again is focussed on functional safety.

5. Conclusions

This work is being undertaken based on clear findings from the first phase of the project. There is strong support for it from a range of areas in the industry, and the outputs are being targeted at an identified need for the industry. The work is funded to continue to completion and further outputs will be presented as the work progresses. This is currently expected to be concluded towards the end of 2013.

References

1. Bearfield, G.J and Short, R, "Standardising safety engineering approaches in the UK railway industry," *6th IET International Conference on System Safety*, 2011, pp.1-5, (2011).
2. CENELEC. *Railway applications – The specification and demonstration of Reliability, Availability, Maintainability and Safety (RAMS),* EN50126:1999 (1999).
3. CENELEC. *Railway applications – Communications, signalling and processing systems – Software for railway control and protection systems*, EN50128:2001 (2001).
4. CENELEC. *Railway applications – Communication, signalling and processing systems – Safety related electronic systems for signalling* EN50129:2003 (2003).
5. CENELEC. *Railway applications. Classification system for railway vehicles. Part 4. Function groups* EN15380-4 (2009).
6. European Commission. Regulation for the CSM on risk assessment (CSM regulation): *Commission regulation (EC) no 352/2009 of 24 April 2009 on the adoption of a common safety method on risk evaluation and assessment* as referred to in Article 6 (3)(a) of Directive 2004/49/EC of the European Parliament and of the Council", The Journal, volume, pp. 110-120, (2000).
7. RSSB, *Engineering Safety Management (Yellow Book*), volumes 1 and 2, (2007).
8. RSSB, *T955: Hazard Analysis and Risk Assessment for Rail Projects, Phase 1: Final Report* (2012).
9. EBA, Sicherheitsrichtlinie Fahrzeug (SIRF) http://www.eba.bund.de/nn_309866/DE/Infothek/Fahrzeuge/Fahrzeugtechnik/funktionaleSicherheit/funktionale__Sicherheit__node.html?__nnn=true (2012)
10. CENELEC. *Railway applications – The specification and demonstration of Reliability, Availability, Maintainability and Safety (RAMS), Part 1: Generic RAMS process. Draft pr* EN50126-1:2012 (2012).

Automatic Construction of a Reliability Model for a Phased Mission System

K.S.Stockwell and S.J.Dunnett
Department of Aeronautical and Automotive Engineering
Loughborough University
Loughborough, Leics. LE11 3TU, UK

Abstract

There are a number of mathematical modelling techniques available to determine the reliability of any system design, for example Fault Trees, Event Trees etc. These models relate the performance of the system to the performance of the components of which it is comprised and are generally quite difficult to construct. Once constructed the analysis of the models can be performed using commercially available software. This stage of the analysis is well developed and can be performed efficiently. The model construction however is a lengthy process and reduces the impact of the reliability study on the system design. One way of improving this situation would be to automate the construction process. In this work a procedure is developed to automatically generate a reliability model, based upon Petri Nets, for a system undertaking a phased mission.

1. Introduction

The design stage of any new system is a critical time to ensure that the system meets all required standards. The reliability of the system must be determined to ensure the standards are met. If this can be done alongside the system design in an efficient manner alternative design solutions can be investigated and the direction of the design could be influenced. A number of mathematical modelling techniques exist to determine the reliability of a system, such as fault trees, event trees and Markov analysis etc. These models require not only a detailed understanding of the system design but also understanding of the techniques themselves and hence a specialist group or team is often brought into model the reliability of system designs. This leads to a lack of project cohesion and as the development of the models generally takes a significant amount of time the results obtained are often too late to effectively influence the system design.

Over the years much work has been performed on the analysis of the models once constructed and this is now well developed and can be performed quickly. The area that still involves significant time and effort is the construction of the models. One way of improving this situation is to automate the process, thus enabling the reliability assessment to be performed alongside the design ensuring full use is made of the results. In the past the automatic construction of fault trees has received the most attention in this area. The most commonly adopted approaches include digraphs (Lapp & Powers 1977), decision tables (Salem et al 1977), transition tables (Taylor 1982) and mini fault trees (Kelly & Lees 1986). All these approaches have some form of restriction on their application and so no one method can be applied to all systems. Despite the restrictions on the use of fault trees to

model systems reliability the automation of the other modelling techniques has received little attention in the past. The aim of the work presented here is to outline an approach to automate the generation of a reliability model for a system undertaking a phased mission. Non-repairable and repairable systems have been considered.

A phased mission is defined as a sequence of tasks (phases) which must be completed to achieve the mission objective. For each task to be completed a different sub-set of the system's capabilities need to function. Clearly, the causes of failure will also be different in each phase. For the mission to be completed successfully all of the phases must have been completed successfully. The main techniques that have been used in solving phased mission problems are Fault Tree Analysis, Markov Analysis and simulation. Both Fault Trees and Markov suffer from the issue that the models become very large for such problems and this increases with the complexity of the problem and the number of phases in the mission. Simulation techniques however are well suited to modelling such situations as their computational nature allows for complex scenarios to be considered. One such technique that allows for simple graphical representation as well as significant modelling power is the Petri net.

2. Petri Nets

Petri nets are a graph based tool that can be used to model the dynamics of many types of system, see Schneeweis (1999). Specifically, a Petri net is a directed bipartite graph in which each node represents either a transition or place, shown in diagrams as a bar or hollow circle respectively. Directed arcs linking places to transitions are known as inputs and those connecting transitions to places are known as outputs. In addition, multiple input or output arcs can link the same place and same transition, with the number of arcs known as the multiplicity, often represented as a single arc with a backslash through it and a positive integer denoting the multiplicity. If there is no backslash then the multiplicity is one. Places may contain 0 or more tokens, represented by filled circles, and it is the distribution of tokens through the net, known as the net marking, that determines the state of the system. Each transition is associated and labelled with a time delay which may be fixed or determined from a distribution. When the number of tokens in a place matches or exceeds the number of input arcs, the transition is enabled and will fire once it has remained enabled for the duration of its associated delay, in which case the tokens are consumed from the input places, and deposited in the output places - thus altering the marking of the net and therefore the state of the system. The number of tokens consumed from the input place is equal to the number of input arcs and the number of token deposited in the output place is equal to the number of output arcs. If the marking of the net changes and disables a previously enabled transition, then that transition and its delay duration are reset. Only one transition can occur at any instant of time, regardless of the number of transitions that are enabled. An example of a transition, showing the before and after net markings, is shown in Figure 1.

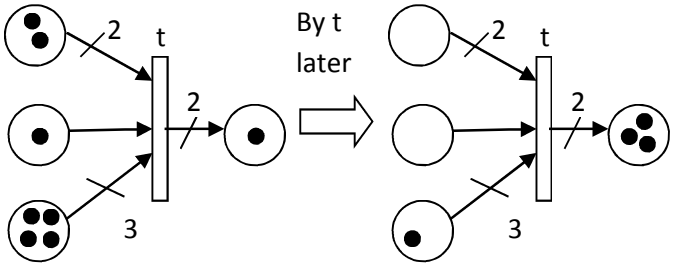


Figure 1. Transition enabling and firing

The figure shows a transition with 3 input places with a multiplicity of 2, 1 and 3 from top to bottom. In the net on the left of the figure the transition is enabled as each input place contains at least as many tokens as its input arcs. Hence after the time delay associated with the transition, t, it fires. Tokens, equal to the number of input arcs, are taken from each input place and tokens equal to the number of output arcs, 2, are deposited into the output place. This is shown in the net on the right of the figure.

3. Automated Model Generation

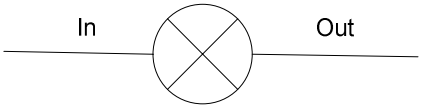
The aim of the procedure outlined in this work is to take a description of a system and its mission in different phases and to generate a petri net model that will determine the reliability of that system.

3.1 System and mission description

The first stage of the procedure is to input into the system all information that is required to construct and run the model. These can be broken down into the following categories:

1. Component description in the form of component models including the component failure modes,
2. System structure in the form of a system topology diagram,
3. Mission Description in the form of phase models and, initial and starting conditions,
4. System failure modes.
5. Failure and repair data.
6. Maintenance strategies

The component models are in the form of decision tables which describe how the component reacts to inputs from other components in the system, depending on the current state of the component. For example, a bulb which has the 2 states of working (W) and failed (F) would have the decision table shown in Table 1. In the table the inputs (In) and outputs (Out) to the bulb would be connected to other components in the system. In the table, C and NC denote current and no current and the ‘-‘ entries are don’t care states where the state or input is irrelevant to the output.



In	State	Out
C	W	C
NC	-	NC
-	F	NC

Table 1. Decision Table for a bulb in a circuit

Operational mode tables, similar to the state transition tables used by (Majdara & Wakabayashi 2009), are also adopted to model components with different operating modes. These tables describe how the mode of operation can be changed, when a command to the component is introduced. For example, if a switch, which is currently open is commanded by an operator to close, as long as the switch is in a working condition, the switch would change mode from open to closed. The operational mode table for a switch is shown below, where mode 1 is the current mode, In is the input from the operator and mode 2 is the resulting switch mode:

Mode 1	In1	State	Mode 2
Closed	-	FCL	Closed
Closed	CL	-	Closed
Closed	OP	W	Open
Closed	NA	-	Closed
Open	-	FOP	Open
Open	CL	W	Closed
Open	OP	-	Open
Open	NA	-	Open

Table 2 Operational model table for a switch

In the table CL, OP, NA, FCL and FOP denote closed, open, no action, failed closed and failed open respectively.

The system topology diagram describes how the components are linked together. The phase models describe, in the form of a phase transition table, the different phases the mission can enter with the system condition needed to transition from one phase to another. The initial conditions are the conditions the components must satisfy in order for the mission to commence. The failure and repair data is necessary for each component in the system to determine a reliability estimate. The maintenance strategies are the different strategies applied to the components within the system, for example preventative maintenance.

3.2 Model Construction

From the information given, as described in section 3.1, software has been developed that will generate a Petri Net model. The model is made up of different Petri Net types that connect to each other, these types are: Component PN (CPN), Component Model PN (CMPN), Circuit PN (CIPN), System PN (SPN) and Phase PN (PPN).

The component PN's are simple nets generated from the component description. They link the working and failed states for components and

incorporate the appropriate maintenance strategy. An example of a component PN for a component whose failure is revealed is shown below. The net includes places to record when maintenance is, or is not, taking place hence allowing for maintenance resources to be built into the net and also downtime recorded.

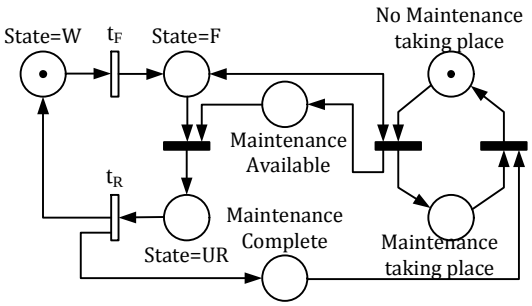


Figure 2. Component Petri net for a component with a revealed failure

For some components the failed state will be dependent upon the current operation mode, an example of such a CPN is shown in Figure 3 for a switch.

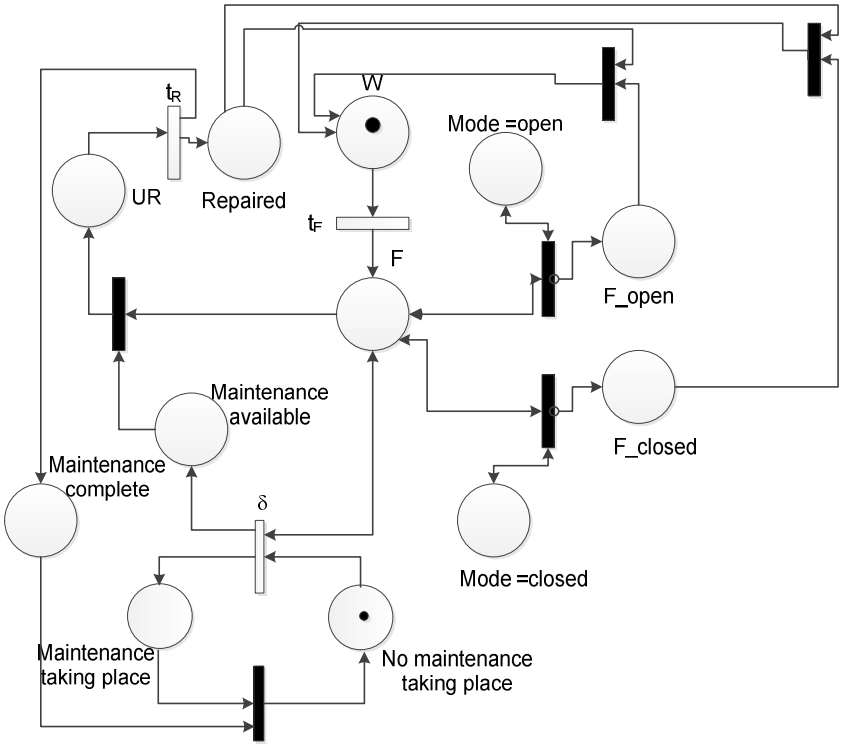


Figure 3. Component Petri net for a switch

The Component Model PN's use the decision tables, operational mode tables and the system topology diagram to generate nets that connect the inputs, states and outputs for the components. For example, considering the decision table for a bulb shown in Table 1 the CMPN is shown below. The places representing the component working and failing, W and F, are connected to the CPN, hence the arrows in the figure.

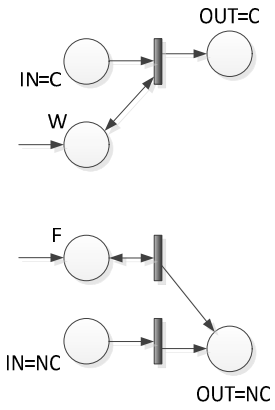


Figure 4. Component model Petri Net for a bulb

The Circuit PNs represent the flow of current in any electrical circuits that are present and are used to identify, given the state or mode of the components in the system, whether there is current, or not, in the circuit.

The system PN links the CMPNs into the system structure and the Phase PN describes the mission undertaken by the system, created using the phase transition table.

These different nets interact with each other: The CMPN's link together to form the SPN. The CPN's pass information to the SPN and the CIPN regarding the states of the components. The SPN pass the information about the current operating modes of the appropriate component to the CPN's and the CIPN's. The PPN moves through the different phases of the mission and obtains the information regarding the phase transition requirements from the SPN. If the system makes a transition to a failed state this will be represented by a phase within the PPN and hence failure and reliability information will be output from the PPN.

3.2 Model simulation

The PN's described above can be used to simulate a systems reliability. Software has been developed to perform this task. The steps undertaken are:

1. From the initial conditions of the system tokens are placed in the relevant places within the CPN's.
2. Place a token in the place representing the first phase in the PPN.
3. Randomly sample failure and repair times for the components from the relevant distribution.
4. Search through each of the immediate transitions in the CPN, SPN and PPN and determine if any are enabled, if so fire them.
5. If the operating mode of a component changes then check the CIPN's to determine if the circuits within the system, if there are any, are passing, or not passing, current. This may affect places within the SPN.
6. Repeat step 4.

When any transition is fired test if any of the following condition are satisfied:

a) a phase transition condition is satisfied. If mission has finished, failure or success, log results and start new simulation.

b) In phase conditions are satisfied for current phase. Check for next timed transition and fire.

3.3 Software

The general structure of the software is shown in figure 5

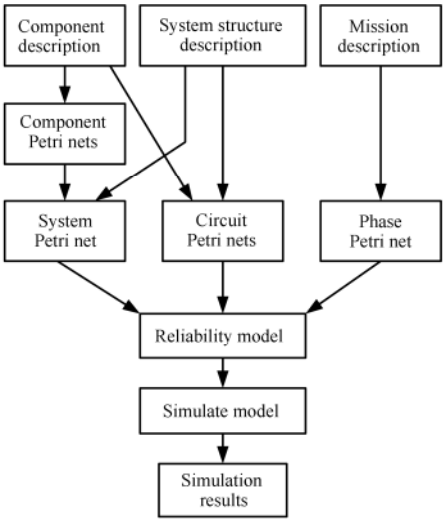


Figure 5. Software structure diagram

The information regarding the components: failure information, repair information, maintenance strategy, etc. is entered by the system designers through a text file, alongside the system topology to be modelled. Using this information the software generates the decision and operational mode tables for all the components in the system. The component information is used to generate the CPN's and the tables are used to generate the CMPN's and hence the SPN. The component failure and repair information is used to generate the times to failure, t_F, and times to repair, t_R, that are the time delays for the transitions in the CPN's. Any circuits in the system are identified by the software by locating all components that are able to pass current and that exist within a path, containing a power supply, that start and end with the same component. Once all the circuits are identified the software automatically generates the CIPN's.

The description of the mission considered is also input by the designers and the software generates a phase transition table that details the different phases and the system conditions for transition between them. The software then uses this table to generate the PPN.

4. Example

The procedure and software have been applied to a non-repairable pressure tank system, see Stockwell and Dunnett (2013) for more detail. For this non-repairable system the procedure was found to be efficient and to give reliable results. In this current work the procedure has been extended to consider repairable systems. In order to demonstrate the procedure a simple bulb system will be described in detail.

The aim of the system is to light the bulb for a fixed period of time. Initially the system is as shown in Figure 7 with the manual switch open. Power supply 2

(PS2) is a cold standby for power suppy 1 (PS1). For ease of presentation here it is assumed that the switching mecahnism between the supplies is perfectly reliabile. The operator closes the switch and the bulb comes on, after a time T_1 hours the operator opens the switch and the bulb goes off. It is assumed that all failures are revealed and if engineers are available maintenance will start immediately.

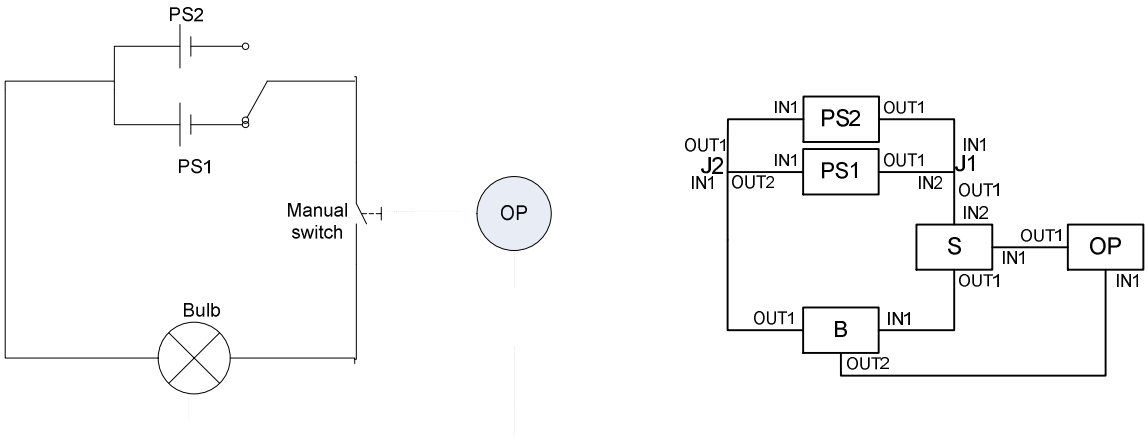


Figure 6 Simple bulb system and associated topology diagram

The topology diagram for the system and the decision and operating mode tables for the components are shown in Figure 6 and tables 3-7.

Time	In1	State	Out1
t=0	OFF	W	CL
$0<t\leq T_1$	OFF	W	NA
$0<t<T_1$	ON	W	NA
$t=T_1$	ON	W	OP
-	-	F	NA

Table 3. Decision table for operator

In1	State	Out1	Out2
C	W	C	ON
NC	-	NC	OFF
-	F	NC	OFF

Table 4. Decision table for bulb.

In2	State	Out1
C	CL	C
-	OP	NC
NC	-	NC

Table 5. Decision table for switch.

In1	State	Out1
C	W	C
-	F	NC
NC	-	NC

Table 6. Decision table for power supply

In1	In2	Out1	Out2
C	-	C	
-	C	C	
NC	NC	NC	
C		C	C
NC		NC	NC

Table 7. Decision table for junction 1 and 2.

In the case of the operator, the output will depend upon time, this dependence is contained in the phase descriptions and is accounted for in the decision table by adding an extra column for time. The manual switch is the only component considered that has more than one operational mode and hence it also has an operating mode table as given in table 2.

The failure modes of the system have been broken down into ‘failure to start up’, ‘failure to keep bulb alight for T_1 hours’, ‘failure to turn bulb off’. In the automation procedure the system failure states have been modelled as separate states and hence there are 6 phases to consider:

Phase 1: Discrete Phase: System start-up
Phase 2: Bulb on for duration of T_1 hours
Phase 3: Fail to start up
Phase 4: System fails to keep bulb lit
Phase 5: System fails to turn off bulb
Phase 6: Mission Success

Phases 3-5 are failure states.
From the description the phase transition table can be determined.

Time	From Phase	To Phase	Condition
0	1	2	B out2=ON
δ	1	3	B out2=OFF
T_1	2	6	S1 mode=open
-	2	4	B out2=OFF
T_1	2	5	S1 mode=closed

Table 8 Phase transition table.

4.1 Petri Nets

As all components are revealed failures the CPN’s will be of the form shown in figure 2. The loop containing the places ‘Maintenance taking place’ and ‘No maintenance taking place’ will be one loop that feeds into all components and will contain the information of whether engineers are available to undertake the maintenance.

The nets for PS1 and PS2 are adapted to account for the fact that the components are in cold standby, see Figure 7. The part of the net inside the large dashed lines corresponds to PS1 and the part inside the dots, PS2. The dashed arrows from the component states feed into the CMPN’s.

The information on the extent and level of redundancy in a system would be included in the detail input by the system designers. A library of CPN’s is being built up for all situations.

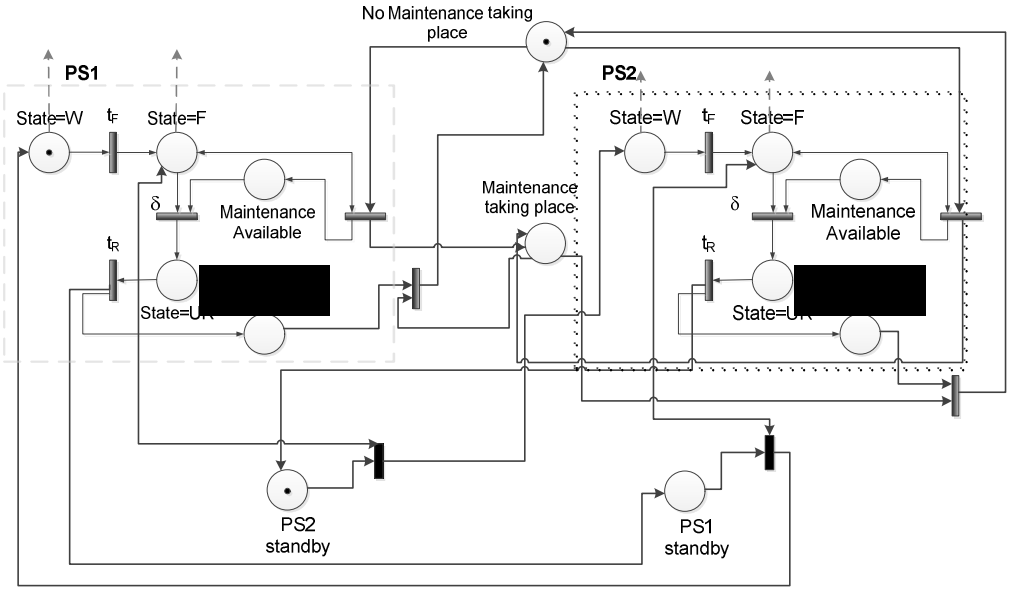


Figure 7 . CPN's for PS1 and PS2.

The CMPN's were generated from the tables 3-8 above and linked together using the system topology diagram to form the SPN which is shown in figure 8. The separate CMPN's can be identified in the figure by the labels , OP, S, B, J2, PS1, PS2, J1 corresponding to Operator, switch, bulb, junction 2, power supply 1, power supply 2 and junction 1 respectively. The dotted arrows in the figure that input into the components states are passing the information from the CPN's.

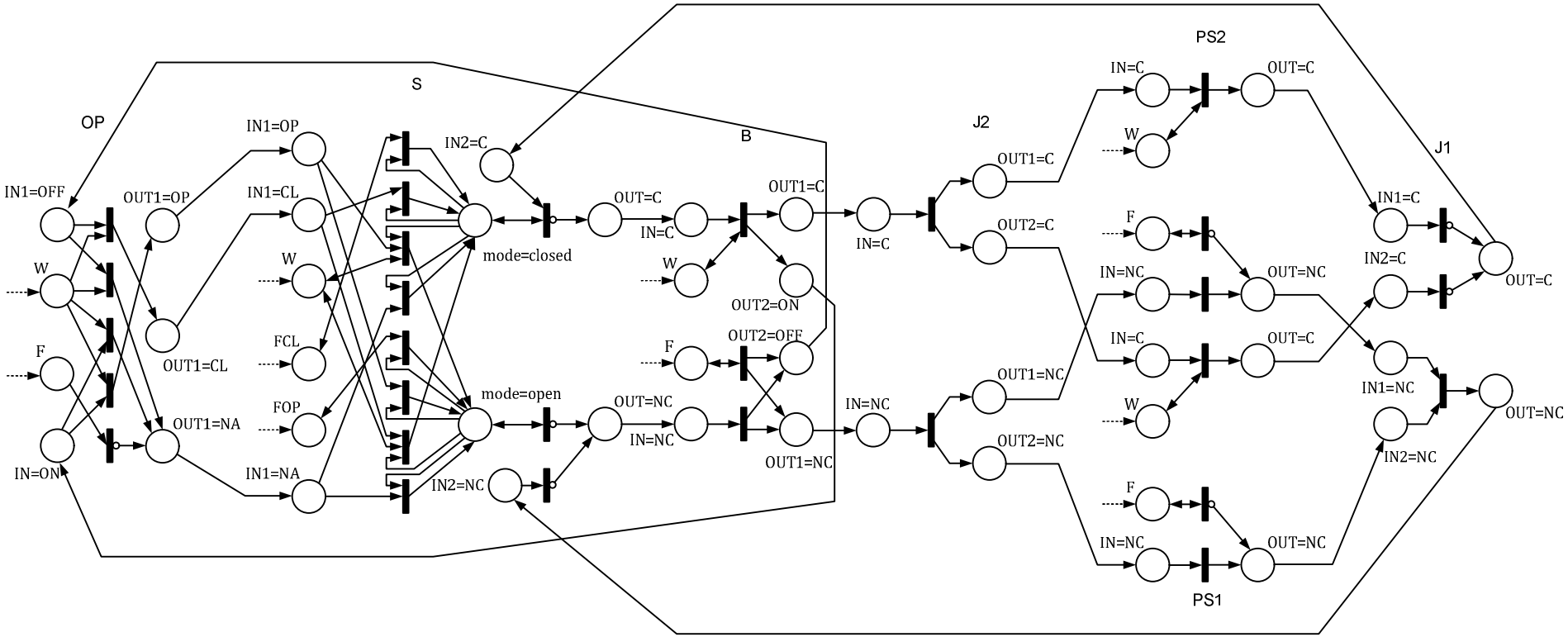


Figure 8. SPN for the bulb system

In this example 2 circuits are identified, one containing PS1 and one containing PS2. The CIPN's for the 2 circuits are the same with only the place for the power supply changing. The CIPN for circuit 1 is shown in figure 9. The place for 'current in circuit 1' will output current to all appropriate inputs to the components in the circuit in the SPN.

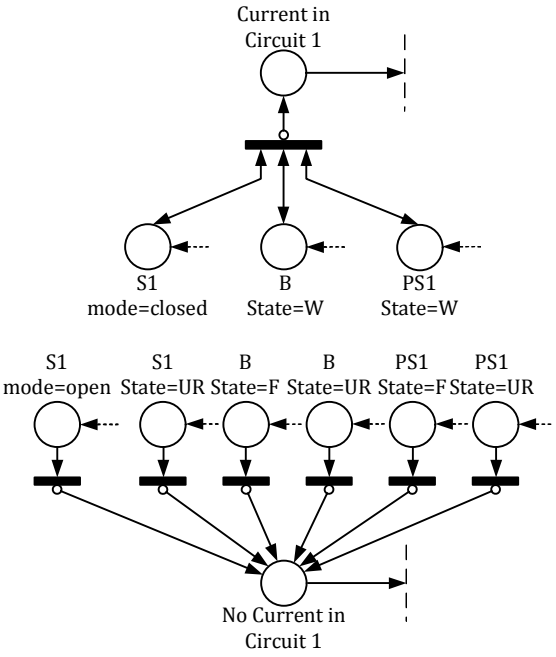


Figure 9. CIPN for one of the circuits in the system

The PPN generated from the phase transition table is shown in figure 10. The dotted arrows in this figure connect the places to places in the SPN.

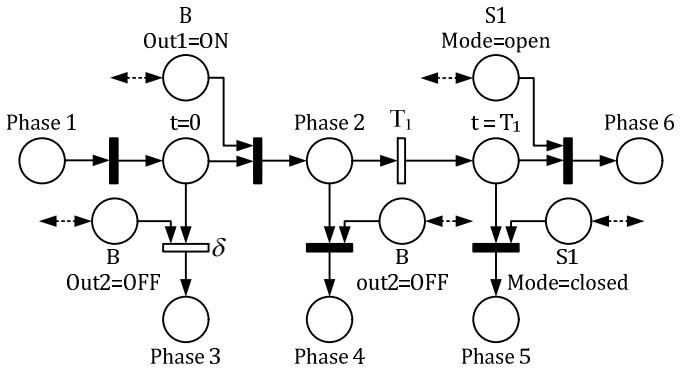


Figure 10. PPN for the mission

Having generated the PN's the steps outlined in section 3.2 are then undertaken to determine the reliability of the system design.

5. *Conclusions*

In the work presented here a procedure for automatically generating a reliability model based on Petri nets and simulation, from a description of a system and its operation has been described. The procedure has been demonstrated by applying it to a simple repairable example. The procedure is currently been applied to a more complex repairable system in order to validate its general use.

In the future in order to improve this method of automation, it is hoped that a technique taking the system description in the form of a computer aided design (CAD) diagram or a piping and instrumentation diagram (P&ID) and then generating the component tables will be developed.

References.

Kelly, B.B. & Lees, F.P. (1986). The propagation of faults in process plants, *Reliability Engineering & System Safety*, 16, 1-108.

Lapp, S.A. & Powers, G.J. (1977). Computer-aided synthesis of fault trees. *IEEE Transactions on Reliability,* R-26, 2-13*.*

Majadera, A & Wakabayashi, T. (2009). Component based modelling of systems for automated fault tree generation, *IEEE Transactions on Reliability* 94: 1076-1086

Mura, I. & Bondavalli, A. (2001). Markov regenerative stochastic Petri nets to model and evaluate phased mission systems dependability. *IEEE Transactions on Computers* 50(12): 1337-1351

Salem, S.L., Apostolakis, G.E. & Okrent, D. (1977). A new methodology for the computer-aided construction of fault trees. *Annals of Nuclear Energy* 4: 417-433

Schneeweis, W.G. (1999). *Petri Nets for Reliability Modeling.* LiLoLe-Verlag GmbH*,*

Stockwell, K.S. and Dunnett, S.J. (2013) Application of a Reliability Model Generator to a Pressure Tank System, *Int. J.Automation and Computing,* 10, 9-17.

Taylor, J.R. (1982). An algorithm for fault tree construction. *IEEE Transactions on Reliability* R-31(2): 137-146

Recent Advances in System Reliability using the Survival Signature

Frank P.A. Coolen[a,1], **Tahani Coolen-Maturi**[b]
Abdullah H. Al-nefaiee[a], **Ahmad M. Aboalkhair**[c]

[a]Department of Mathematical Sciences, Durham University, UK.
[b]Durham University Business School, Durham University, UK.
[c]Department of Applied Statistics and Insurance, Mansoura University, Egypt.

Abstract

The theory of system signatures provides a powerful framework for reliability assessment for systems consisting of exchangeable components. However, its generalization to systems with multiple types of components is very complicated, if not impossible. Recently, we have introduced the concept of 'survival signatures' as an attractive and more powerful alternative. For systems with single types of components, the survival signatures are closely related to the signatures. However, the generalization to systems with multiple types of components is conceptually straightforward for survival signatures.

In this paper, we give an introductory overview of survival signatures, explaining their use in reliability quantification. In addition, we present how survival signatures of subsystems can be combined to derive a system's survival signature, and we show how survival signatures change in case of replacement of a component. Finally, we briefly discuss some related research challenges.

1. Introduction

In recent decades, system signatures have proven to be a powerful tool for qualifying reliability of coherent systems consisting of components with random failure times that are independent and identically distributed (*iid*), although this assumption can be relaxed to assuming exchangeability. The system signature can be used to quantify aspects of system reliability such as its failure time distribution. An attractive feature of describing system structures through signatures is the possibility to compare the reliability of different systems based on stochastic ordering of their signatures, as long as the components' failure times in these systems are all *iid* (or exchangeable). A detailed introduction and overview to system signatures is presented by Samaniego [12], some recent advances are reviewed by Eryilmaz [8].

Consider a system consisting of m components with *iid* failure times. Throughout this paper it is assumed that the system is coherent, but the approach can also be generalized to non-coherent systems. Let the random failure time of the system be T_S, and let $T_{j:m}$ be the j-th order statistic of the m random component

[1]Corresponding author: frank.coolen@durham.ac.uk

failure times for $j = 1, \ldots, m$, with $T_{1:m} \leq T_{2:m} \leq \ldots \leq T_{m:m}$. The system's signature is the m-vector q with j-th component

$$q_j = P(T_S = T_{j:m}) \tag{1}$$

so q_j is the probability that the system failure occurs at the moment of the j-th component failure. It is natural to assume that $\sum_{j=1}^{m} q_j = 1$, so the system functions if all components function, has failed if all components have failed, and system failure can only occur at times of component failures. The survival function of the system failure time can be derived by

$$P(T_S > t) = \sum_{j=1}^{m} q_j P(T_{j:m} > t) \tag{2}$$

If the components' failure times are *iid* with known cumulative distribution function $F(t)$, then

$$P(T_{j:m} > t) = \sum_{r=m-j+1}^{m} \binom{m}{r} [1 - F(t)]^r [F(t)]^{m-r} \tag{3}$$

The essential property of the system signature is clear from Equation (2), namely it enables information of the system structure to be fully taken into account through the signature, and this is separated from information about the random failure times of the components. The main disadvantage of system signatures, however, is that it is effectively impossible to keep this separation when generalizing the concept to systems with multiple types of components, which is crucial for a practically applicable theory as most real-world systems consist of more than a single type of components. Such a generalization would always require probabilities for orderings of order statistics from different probability distributions, corresponding to the different types of components, which would be very difficult to implement. This is explained in detail by Coolen and Coolen-Maturi [5], who introduced the concept of 'survival signature' as an attractive alternative: for systems with just one type of components, the survival signature is closely related to the system signature, but the survival signature can straightforwardly be generalized to systems with multiple types of components.

In Section 2 of this paper we present a short introductory overview of the survival signature. In Section 3 we present how survival signatures of subsystems can be combined to derive a system's survival signature. In Section 4 we show how survival signatures change in case of replacement of a component. Section 5 concludes the paper with a brief discussion of some related research challenges.

2. The survival signature

For a system with m components, we define the state vector $\underline{x} = (x_1, x_2, \ldots, x_m) \in \{0, 1\}^m$ with $x_i = 1$ if the ith component functions and $x_i = 0$ if not. The labelling

of the components is arbitrary but must be fixed to define $\underline{x}$. The structure function $\phi : \{0,1\}^m \to \{0,1\}$, defined for all possible $\underline{x}$, takes the value 1 if the system functions and 0 if the system does not function for state vector $\underline{x}$. Actually, this definition of the structure function can be generalized by defining it as the probability that the system functions for state vector $\underline{x}$. We do not consider this further here, but it may be of practical relevance, for example if system functioning is defined as the system meeting a certain demand where this demand might be random. The survival signature approach can also be developed for this generalization. We restrict attention to coherent systems, which means that $\phi(\underline{x})$ is not decreasing in any of the components of $\underline{x}$, so system functioning cannot be improved by worse performance of one or more of its components. We further assume that $\phi(\underline{0}) = 0$ and $\phi(\underline{1}) = 1$, so the system fails if all its components fail and it functions if all its components function. These last two assumptions could be relaxed but are reasonable for most practical systems, and they simplify the presentation in this paper.

Coolen and Coolen-Maturi [5] introduced the system survival signature as alternative to the system signature. For a system consisting only of components with *iid* failure times (this assumption can be relaxed to exchangeable failure times), the survival signature, denoted by $\Phi(l)$, for $l = 1, \ldots, m$, is defined as the probability that the system functions given that *precisely* l of its components function. For coherent systems, $\Phi(l)$ is an increasing function of l, and with the second assumption above we have $\Phi(0) = 0$ and $\Phi(m) = 1$. There are $\binom{m}{l}$ state vectors $\underline{x}$ with precisely l components $x_i = 1$, so with $\sum_{i=1}^{m} x_i = l$; let S_l denote the set of these state vectors. Due to the *iid* assumption for the failure times of the m components, all these state vectors are equally likely to occur (this also holds under the weaker assumption of exchangeability), hence

$$\Phi(l) = \binom{m}{l}^{-1} \sum_{\underline{x} \in S_l} \phi(\underline{x}) \tag{4}$$

Coolen and Coolen-Maturi [5] called $\Phi(l)$ the survival signature because, by its definition, it is closely related to survival of the system, and it is close in nature to the system signature as will soon be clear.

Let $C_t \in \{0, 1, \ldots, m\}$ denote the number of components in the system that function at time $t > 0$. Given the cumulative distribution function $F(t)$ for the failure times of the components, it is clear that, for $l \in \{0, 1, \ldots, m\}$

$$P(C_t = l) = \binom{m}{l} [F(t)]^{m-l} [1 - F(t)]^l \tag{5}$$

This leads to

$$P(T_S > t) = \sum_{l=0}^{m} \Phi(l) P(C_t = l) \tag{6}$$

It is clear from Equation (6) that the term $\Phi(l)$ takes the structure of the system into account and is separated from the information about the failure time distribution for the components, which is included through the term $P(C_t = l)$.

Hence, the survival signature achieves the same separation of these two aspects which is the main advantage of the system signature, as mentioned in Section 1. This is not surprising, as the survival signature and the system signature are closely related. Indeed, it is easily seen that the following equality holds [5]

$$\Phi(l) = \sum_{j=m-l+1}^{m} q_j \tag{7}$$

Equation (7) is logical when considering that the right-hand side is the probability that the system failure time occurs at the moment of the $(m-l+1)$-th ordered component failure time or later. This is exactly the moment at which the number of functioning components in the system decreases from l to $l-1$, hence the system would have functioned with l components functioning. Due to this direct relation between the survival signature and the system signature, for systems consisting of one type of components with *iid* failure times, it is clear that the analytic possibilities provided by the signature are also provided by the survival signature. For example, methods for comparison of two systems based on the survival signatures were presented by Coolen and Coolen-Maturi [5]. However, when we are considering systems with multiple types of components, for which the system signature cannot be generalized, the survival signature becomes far more attractive.

Let us consider a system with $K \geq 2$ types of components, with m_k components of type $k \in \{1, 2, \ldots, K\}$ and $\sum_{k=1}^{K} m_k = m$. Assume that the random failure times of components of the same type are *iid*, while full independence is assumed for the random failure times of components of different types. Due to the arbitrary ordering of the components in the state vector, components of the same type can be grouped together, leading to a state vector that can be written as $\underline{x} = (\underline{x}^1, \underline{x}^2, \ldots, \underline{x}^K)$, with $\underline{x}^k = (x_1^k, x_2^k, \ldots, x_{m_k}^k)$ the sub-vector representing the states of the components of type k. Let the ordered random failure times of the m_k components of type k be denoted by $T_{j_k:m_k}^k$.

The survival signature for such a system is denoted by $\Phi(l_1, l_2, \ldots, l_K)$, for $l_k = 0, 1, \ldots, m_k$, which is defined to be the probability that a system functions given that *precisely* l_k of its m_k components of type k function, for each $k \in \{1, 2, \ldots, K\}$ [5]. There are $\binom{m_k}{l_k}$ state vectors $\underline{x}^k$ with precisely l_k of its m_k components x_i^k equal to 1, so with $\sum_{i=1}^{m_k} x_i^k = l_k$; let S_l^k denote the set of these state vectors for components of type k. Furthermore, let $S_{l_1,\ldots,l_K}$ denote the set of all state vectors for the whole system for which $\sum_{i=1}^{m_k} x_i^k = l_k$, $k = 1, 2, \ldots, K$. Due to the *iid* assumption for the failure times of the m_k components of type k, all the state vectors $\underline{x}^k \in S_l^k$ are equally likely to occur, hence

$$\Phi(l_1, \ldots, l_K) = \left[\prod_{k=1}^{K} \binom{m_k}{l_k}^{-1} \right] \times \sum_{\underline{x} \in S_{l_1,\ldots,l_K}} \phi(\underline{x}) \tag{8}$$

Let $C_t^k \in \{0, 1, \ldots, m_k\}$ denote the number of components of type k in the system that function at time $t > 0$. If the probability distribution for the failure time of

components of type k is known and has CDF $F_k(t)$, then for $l_k \in \{0, 1, \ldots, m_k\}$, $k = 1, 2, \ldots, K$,

$$\begin{aligned} P(\bigcap_{k=1,\ldots,K} \{C_t^k = l_k\}) &= \prod_{k=1}^{K} P(C_t^k = l_k) \\ &= \prod_{k=1}^{K} \binom{m_k}{l_k} [F_k(t)]^{m_k - l_k} [1 - F_k(t)]^{l_k} \end{aligned} \tag{9}$$

The probability that the system functions at time $t > 0$ is

$$P(T_S > t) = \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(l_1, \ldots, l_K) P(\bigcap_{k=1}^{K} \{C_t^k = l_k\}) \tag{10}$$

The survival signature $\Phi(l_1, \ldots, l_K)$ must be derived for all $\prod_{k=1}^{K}(m_k + 1)$ different $(l_1, \ldots, l_K)$. This information must anyhow be distracted from the system if one wishes to assess its reliability. The survival signature only has to be calculated once for any system, similar to the (survival) signature for systems with a single type of components. The main advantage of (10) is that again the information about system structure is fully separated from the information about the components' failure times, and the inclusion of the failure time distributions is straightforward due to the assumed independence of failure times of components of different types.

The survival signature $\Phi(l_1, \ldots, l_K)$, as introduced for $K \geq 1$ different types of components, can be used as long as the failure times of components of the same type are exchangeable. In this paper we have actually made the stronger assumption of *iid* failure times as this allows $P(C_t^k = l_k)$ to be written in terms of $F_k(t)$, as shown in Equation (9). Furthermore, we assumed that the failure times of components of different types are fully independent, which justifies the first equality in Equation (9). The main idea of the survival signature, however, can be applied without these assumptions, in which case the joint probability at the left-hand side of Equation (9) must be specified.

Example 1.
Coolen and Coolen-Maturi [5] presented a small example involving the system with $K = 2$ types of components, types 1 and 2, as presented in Figure 1. The survival signature for this system is presented in Table 1, for further explanation and illustration of a the system survival function for specific probability distributions for the two types of components we refer to [5]. We will return to this example in Section 4.

3. Combining survival signatures of subsystems

Computation of the survival signature is complicated for systems of realistic size. Gaofeng, Zheng and Taizhong [11] showed how the system signature can be derived from the signatures of two subsystems, if the system consists of these two

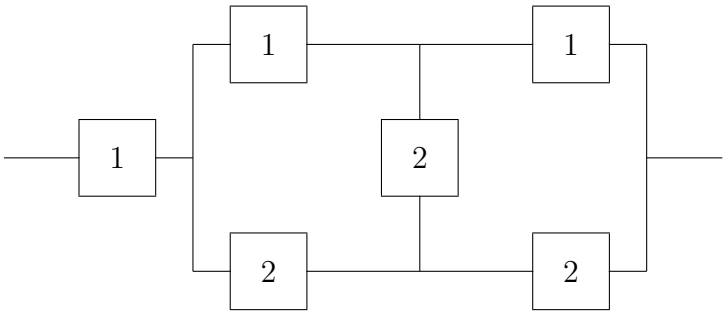


Figure 1: System with 2 types of components (Ex. 1)

l_1	l_2	$\Phi(l_1, l_2)$	l_1	l_2	$\Phi(l_1, l_2)$
0	0	0	2	0	0
0	1	0	2	1	0
0	2	0	2	2	4/9
0	3	0	2	3	6/9
1	0	0	3	0	1
1	1	0	3	1	1
1	2	1/9	3	2	1
1	3	3/9	3	3	1

Table 1: Survival signature of the system in Figure 1

subsystems in either series or parallel configuration. Repeated application of their method enables quite straightforward computation of the signature of any system consisting of any number of subsystems such that the overall structure can be created through a sequence of series or parallel configurations. In this section we present a similar method for the survival function of a system consisting of two subsystems in either series or parallel configuration; by repeated use this enables the survival signatures for quite a substantial range of systems to be computed relatively easily.

Suppose that a system consists of 2 subsystems for which the survival signatures are known. Let the system consist of $K \geq 1$ types of components, with m_k components of type k, for $k = 1, \ldots, K$, of which $m_k^r \geq 0$ are in subsystem r, for $r = 1, 2$. Let subsystem r consist in total of m^r components, so $m^r = \sum_{k=1}^{K} m_k^r$. We denote the survival signature for subsystem r by $\Phi^r(l_1^r, l_2^r, \ldots, l_K^r)$, for $l_k^r = 0, 1, \ldots, m_k^r$. For ease of notation, we define $\Phi^r(l_1^r, l_2^r, \ldots, l_K^r) = 0$ if $l_k^r > m_k^r$ for any $k \in \{1, \ldots, K\}$. If the two subsystems are in series configuration, then the survival signature of the system can be derived, for $0 \leq l_k \leq m_k$, $k \in \{1, \ldots, K\}$,

by

$$\Phi_S(l_1,\ldots,l_K) = \sum_{l_1^1=0}^{l_1}\cdots\sum_{l_K^1=0}^{l_K}\Bigg[\Phi^1(l_1^1,\ldots,l_K^1)\Phi^2(l_1-l_1^1,\ldots,l_K-l_K^1)\times \prod_{k=1}^{K}\binom{m_k^1}{l_k^1}\binom{m_k^2}{l_k-l_k^1}\binom{m_k}{l_k}^{-1}\Bigg] \tag{11}$$

Similarly, if the two subsystems are in parallel configuration, then the survival signature of the system can be derived, for $0 \leq l_k \leq m_k$, $k \in \{1,\ldots,K\}$, by

$$\Phi_P(l_1,\ldots,l_K) = \sum_{l_1^1=0}^{l_1}\cdots\sum_{l_K^1=0}^{l_K}\Bigg[\{1-(1-\Phi^1(l_1^1,\ldots,l_K^1))(1-\Phi^2(l_1-l_1^1,\ldots,l_K-l_K^1))\}\times \prod_{k=1}^{K}\binom{m_k^1}{l_k^1}\binom{m_k^2}{l_k-l_k^1}\binom{m_k}{l_k}^{-1}\Bigg] \tag{12}$$

These results follow from straightforward combinatorial arguments, using the hypergeometric distribution for the probability that precisely l_k^1 of the l_k functioning components of type k are among the m_k^1 components of this type in subsystem 1, and the remainder are among the m_k^2 components of this type in subsystem 2.

Example 2.
We illustrate the method presented in this section by calculating the survival signature for the system in Figure 2, which has three subsystems, labelled by A, B and C, and $K = 3$ types of components. The survival signatures of the three subsystems are easily derived and presented in Table 2. This table presents the survival signatures for the subsystems as functions of the numbers of components of each of these three types, even if not all types of components occur in the subsystem, as this is in line with the notation introduced in this section.

We first use Equation (11) to derive the survival signature for the subsystem which consists of subsystems A and B in series configuration, we refer to this as subsystem AB. This leads to the survival signature presented in Table 3, where apart from $\Phi^{ab}(0,0,0) = 0$ all not presented $\Phi^{ab}(l_1^{ab}, l_2^{ab}, l_3^{ab})$ with $l_1^{ab} \in \{0,1\}$, $l_2^{ab} \in \{0,1,2,3\}$ and $l_3^{ab} \in \{0,1,2\}$ are equal to 1.

We calculate the survival signature of the entire system by combining the survival signature Φ^{ab} for subsystem AB with the survival signature Φ^c for subsystem C, using Equation (12). This leads to the system's survival signature presented in Table 4, where apart from $\Phi(0,0,0) = 0$ all not presented $\Phi(l_1,l_2,l_3)$ with $l_1 \in \{0,1\}$, $l_2 \in \{0,1,2,3,4\}$ and $l_3 \in \{0,1,2,3,4\}$ are equal to 1. To provide more insight into the survival signature, the values in this table are given as fractions that correspond to the number of combinations considered when deriving these values directly.

While direct calculation of the survival signature of this system is still possible, the use of the method presented in this section simplifies it considerably.

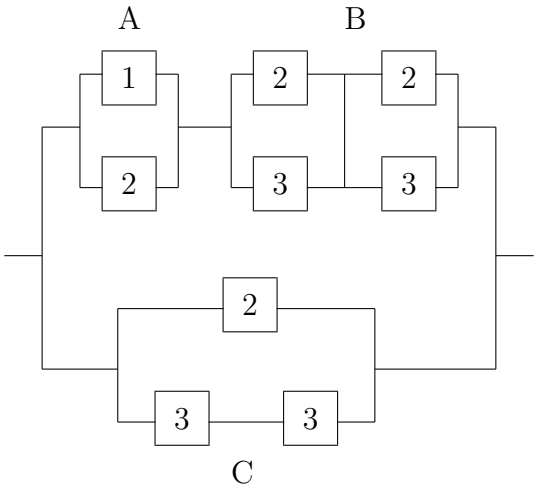


Figure 2: System with 3 types of components (Ex. 2)

Subsystem A				Subsystem B				Subsystem C			
l_1^a	l_2^a	l_3^a	$\Phi^a(l_1^a, l_2^a, l_3^a)$	l_1^b	l_2^b	l_3^b	$\Phi^b(l_1^b, l_2^b, l_3^b)$	l_1^c	l_2^c	l_3^c	$\Phi^c(l_1^c, l_2^c, l_3^c)$
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	1	0	0	0	1	0
0	1	0	1	0	0	2	1	0	0	2	1
1	1	0	1	0	1	0	0	0	1	0	1
				0	1	1	1/2	0	1	1	1
				0	1	2	1	0	1	2	1
				0	2	0	1				
				0	2	1	1				
				0	2	2	1				

Table 2: Survival signatures of subsystems A, B and C in Figure 2

When considering systems with more (types of) components and more subsystems, direct calculation soon becomes impossible while the presented method remains straightforward to implement and leads to fast calculation of the system's survival signature, as long as the system consists of subsystems which can all be combined by sequentially combining two subsystems in series or parallel configuration.

4. The effect of component replacement

In many situations where reliability of systems is of interest, activities such as maintenance or replacement of components are important. If one component of a specific type k is replaced, the failure time distribution of the new component will typically differ from those components in the system that were of the same type before this replacement (unless the component's failure time is assumed to be Exponentially distributed, in which case the replaced component might still be considered to be of the same type as before). So, such an activity will, effectively, imply that there is a new type of component in the system, say type $K + 1$,

l_1^{ab}	l_2^{ab}	l_3^{ab}	$\Phi^{ab}(l_1^{ab}, l_2^{ab}, l_3^{ab})$	l_1^{ab}	l_2^{ab}	l_3^{ab}	$\Phi(l_1^{ab}, l_2^{ab}, l_3^{ab})$
0	0	1	0	0	2	2	2/3
0	0	2	0	1	0	0	0
0	1	0	0	1	0	1	0
0	1	1	0	1	1	0	0
0	1	2	1/3	1	1	1	1/3
0	2	0	0	1	2	0	1/3
0	2	1	1/3	1	2	1	2/3

Table 3: Survival signature of subsystem AB in Figure 2

l_1	l_2	l_3	$\Phi(l_1, l_2, l_3)$	l_1	l_2	l_3	$\Phi(l_1, l_2, l_3)$
0	0	1	0	0	2	3	22/24
0	0	2	1/6	1	0	0	0
0	0	3	1/2	1	0	1	0
0	1	0	1/4	1	0	2	2/6
0	1	1	1/4	1	1	0	1/4
0	1	2	10/24	1	1	1	6/16
0	1	3	12/16	1	1	2	16/24
0	2	0	3/6	1	2	0	4/6
0	2	1	14/24	1	2	1	18/24
0	2	2	27/36	1	2	2	32/36

Table 4: Survival signature of the system in Figure 2

and that the number of components of its earlier type k has been reduced from m_k to $m_k - 1$. We now consider the effect of such a component replacement on the system's survival signature. We restrict attention to replacement of a single component; this can be generalized, following the same principles, to replacement of multiple components.

Let $\Phi(l_1, \ldots, l_{k-1}, l_k, l_{k+1}, \ldots, l_K)$ be the survival signature for a system with K types of components, with $l_k \in \{0, 1, \ldots, m_k\}$ for $k = 1, \ldots, K$. Now suppose that 1 component of type k is replaced by a component of a new type, say type $K+1$. This may really be a new type of component, or just similar to the one that is being replaced but with a different age, hence at any moment in time the value of its failure time distribution differs from that of the other components in the system. We must calculate the survival signature of this system with $K+1$ types of components, which we denote by $\tilde{\Phi}(l_1 \ldots, l_{k-1}, \tilde{l}_k, l_{k+1}, \ldots, l_K, \tilde{l}_{K+1})$, where the tilde is added to emphasize a change compared to the survival signature of this system before the component was replaced.

The numbers m_j of components of types $j \neq k$ remain the same as before the replacement, and now there are $\tilde{m}_k = m_k - 1$ components of type k and $\tilde{m}_{K+1} = 1$ component of the new type $K+1$. So the new survival signature $\tilde{\Phi}(l_1 \ldots, l_{k-1}, \tilde{l}_k, l_{k+1}, \ldots, l_K, \tilde{l}_{K+1})$ must be specified for $l_j \in \{0, 1, \ldots, m_j\}$ for $j \in \{1, \ldots, k-1, k+1, \ldots, K\}$, $\tilde{l}_k \in \{0, 1, \ldots, \tilde{m}_k\}$ and $\tilde{l}_{K+1} \in \{0, 1\}$. This

specification is simplified by the following relationship,

$$\begin{aligned}&\Phi(l_1,\ldots,l_{k-1},l_k,l_{k+1},\ldots,l_K) = \\ &\frac{l_k}{m_k}\times\tilde{\Phi}(l_1\ldots,l_{k-1},l_k-1,l_{k+1},\ldots,l_K,1)+ \\ &\frac{m_k-l_k}{m_k}\times\tilde{\Phi}(l_1\ldots,l_{k-1},l_k,l_{k+1},\ldots,l_K,0)\end{aligned} \tag{13}$$

The proof Equation (13) is based on the probability that the replaced component would be one of the l_k functioning ones out of the m_k components of type k in the original system, or one of the $m_k - l_k$ non-functioning ones. Hence, if one has the fully specified original survival signature Φ available, the computations required in order to fully specify the new survival signature $\tilde{\Phi}$ can, for example, be restricted to computing the values of $\tilde{\Phi}(l_1\ldots,l_{k-1},l_k-1,l_{k+1},\ldots,l_K,1)$, from which the values of $\tilde{\Phi}(l_1\ldots,l_{k-1},l_k,l_{k+1},\ldots,l_K,0)$ follow by Equation (13), and together these fully specify $\tilde{\Phi}$. While this does require new computations, the overall system structure remains the same, and attention can now be restricted to $l_k - 1$ components of type k, with the replaced component assumed to be functioning. The computations to derive $\tilde{\Phi}$ are similar to those for Φ, with a slight simplification due to the assumption that the new component functions with certainty.

Example 3.
To illustrate the effect of component replacement on a system's survival signature, as discussed in this section, consider again the system discussed in Example 1, as presented in Figure 1 with the survival signature given in Table 1. The numbers of components of types 1 and 2 in the original system are $m_1 = 3$ and $m_2 = 3$. Assume that one component of type 2 in the original system is replaced by a component of type 3, leading to the system in Figure 3. The numbers of components of types 1, 2 and 3 in the system after this component replacement are $m_1 = 3$, $\tilde{m}_2 = 2$ and $\tilde{m}_3 = 1$. The survival signature $\tilde{\Phi}$ for this system, after the component replacement, is given in Table 5. The values of $\tilde{\Phi}$ are quite straightforward to verify. It is also easy to confirm that these values, together with the values of Φ in Table 1, satisfy Equation (13). For example, for $l_1 = 2$ and $l_2 = 2$, this equation becomes

$$\Phi(2,2) = \frac{2}{3}\tilde{\Phi}(2,1,1) + \frac{1}{3}\tilde{\Phi}(2,2,0) = \frac{2}{3}\times\frac{1}{2} + \frac{1}{3}\times\frac{1}{3} = \frac{4}{9}$$

To calculate $\tilde{\Phi}$ in Table 5, we only have to calculate the values of $\tilde{\Phi}(l_1,\tilde{l}_2,1)$, as given in the second column of Table 5. This is easier than the original calculation of Φ in Table 1, because it concerns the same system structure but now with the new component of type 3 in Figure 3 certainly functioning. Given the fully specified survival signature Φ of the original system, the values of $\tilde{\Phi}(l_1,\tilde{l}_2,0)$, as given in the first column of Table 5, follow now easily from Equation (13).

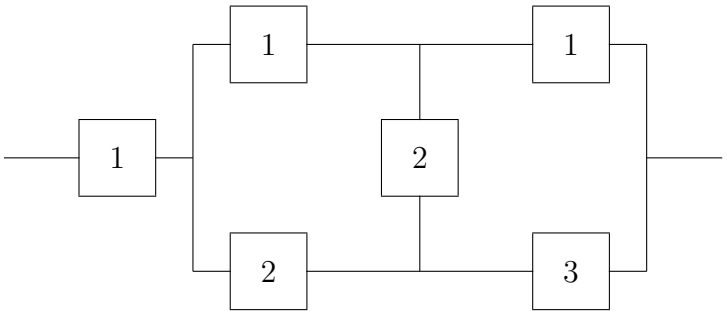


Figure 3: System with 3 types of components (Ex. 3)

l_1	$\tilde{l}_2$	$\tilde{l}_3$	$\tilde{\Phi}(l_1, \tilde{l}_2, \tilde{l}_3)$	l_1	$\tilde{l}_2$	$\tilde{l}_3$	$\tilde{\Phi}(l_1, \tilde{l}_2, \tilde{l}_3)$
0	0	0	0	0	0	1	0
0	1	0	0	0	1	1	0
0	2	0	0	0	2	1	0
1	0	0	0	1	0	1	0
1	1	0	0	1	1	1	1/6
1	2	0	0	1	2	1	1/3
2	0	0	0	2	0	1	0
2	1	0	0	2	1	1	1/2
2	2	0	1/3	2	2	1	2/3
3	0	0	1	3	0	1	1
3	1	0	1	3	1	1	1
3	2	0	1	3	2	1	1

Table 5: Survival signature of the system in Figure 3

5. Concluding remarks

For real-world systems with substantial numbers of components of several types, calculating the survival signature is very computationally demanding. Of course, it only needs to be derived once and then enables quite straightforward quantification of the system reliability for assumed component failure time distributions. If one is interested in a specific reliability criterion, for example whether the system survival time exceeds a certain value with a required probability, then one may not need to calculate the entire survival signature precisely. Partial information about the survival signature, for example if it has been calculated for some but not all $(l_1, \ldots, l_K)$, can straightforwardly be transformed to bounds for the survival signature at other $(l_1, \ldots, l_K)$, using the fact that the survival signature of a coherent system is increasing in every l_k (for non-coherent systems this does not hold). Such bounds for the survival signature can then be used to derive bounds for the system survival function, which can be interpreted as lower and upper

probabilities [6, 7]. If these bounds are already conclusive for a specific overall inference, then it is not necessary to calculate the survival signature more precisely. Similar theory for system signatures is presented by Al-nefaiee and Coolen [1], detailed results for the survival signature will be presented elsewhere.

This paper has not addressed statistical inference for the system survival function, using the survival signature. Due to the separation of information about the system structure, taken into account via the survival signature, and the component failure times, this is relatively straightforward. Nonparametric predictive inference (NPI) methods [3], assuming failure time data for components, has been developed and will be presented elsewhere. This is quite similar to the NPI method for the system survival function using the system signature, for systems with a single type of components [1, 4]. The development of Bayesian methods for such inferences is also an important topic for future research. An interesting further question is whether it is possible to learn about a system's survival signature from failure observations. Recently, Aslett [2] has made interesting contributions to Bayesian learning of the system signature when only data for the whole system are available. This is important for 'black-box' systems, where it is not possible to construct the signature on the basis of available information. It will be interesting to develop a similar approach to learning the system survival signature.

Additional aspects of practical relevance include design decisions for the system structure, for example to include sufficient redundancy to ensure high levels of reliability, and statistical methodology for dealing with competing risks. These topics have been studied within the NPI framework [9, 10], but not yet in relation to system reliability with the use of the survival signature, so these are interesting challenges for research.

There are several established methods to quantify system reliability, for example using fault trees, Bayesian networks or binary decision diagrams. It is important to investigate the possibility to use such methods together with the survival signature, in particular to see if such methods might simplify calculation of the survival signature and enhance applicability of survival signatures. While the emphasis in this paper has been on system reliability, the closely related topic of reliability of networks is of great practical importance, for example in energy provision. Developing the survival signature approach for network reliability is therefore also an important research challenge.

References

[1] Al-nefaiee A.H. and Coolen F.P.A. Nonparametric predictive inference for system failure time based on bounds for the signature. *Journal of Risk and Reliability*, revised version in submission (2013).

[2] Aslett L.J.M. *MCMC for Inference on Phase-type and Masked System Lifetime Models.* PhD Thesis, Trinity College Dublin (*www.louisaslett.com*) (2012).

[3] Coolen F.P.A. Nonparametric predictive inference. In: M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, pp. 968–970 (2011).

[4] Coolen F.P.A. and Al-nefaiee A.H. Nonparametric predictive inference for failure times of systems with exchangeable components. *Journal of Risk and Reliability*, **226** 262–273 (2012).

[5] Coolen F.P.A. and Coolen-Maturi T. On generalizing the signature to systems with multiple types of components. In: W. Zamojski *et al* (Eds.), *Complex Systems and Dependability*, Springer, pp. 115–130 (2012).

[6] Coolen F.P.A., Troffaes M.C. and Augustin T. Imprecise probability. In: M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, pp. 645–648 (2011).

[7] Coolen F.P.A. and Utkin L.V. Imprecise reliability. In: M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, pp. 649–650 (2011).

[8] Eryilmaz S. Review of recent advances in reliability of consecutive k-out-of-n and related systems. *Journal of Risk and Reliability*, **224** 225–237 (2010).

[9] MacPhee I.M., Coolen F.P.A. and Aboalkhair A.M. Nonparametric predictive system reliability with redundancy allocation following component testing. *Journal of Risk and Reliability*, **223** 181–188 (2009).

[10] Maturi T.A., Coolen-Schrijner P. and Coolen F.P.A. Nonparametric predictive inference for competing risks. *Journal of Risk and Reliability*, **224** 11–26 (2010).

[11] Gaofeng D., Zheng B. and Taizhong H. On computing signatures of coherent systems. *Journal of Multivariate Analysis*, **103** 142–150 (2012).

[12] Samaniego F. *System Signatures and their Applications in Engineering Reliability*. Springer (2007).

Degradation Test Analysis: A Case Study

Filippo De Carlo[1], Orlando Borgia[1], Mario Tucci[1]
Industrial Engineering Department, University of Florence, Florence, Italy
Viale Giovan Battista Morgani, 40 - 50134, Firenze, Italy

Abstract
The increasing competition of global markets requires companies to focus their attentions on two main aspects of the internal production system: cost reduction and product quality. Quality is identified with customer satisfaction. This can be converted to the contentment of the customer in terms of attended performance during the entire life cycle of the good. In particular, reliability performance needs to be investigated and tested during the design of the new product. The aim of this study is to evaluate the reliability performance of a model of washing machine for domestic use. The application of degradation and testing analysis was necessary to identify the main deterioration mechanisms. This paper presents an approach to estimate degradation trend of a new machine.
A group of 24 washing machines, divided into three groups of 8 washing machines, has been tested. Increasing loads were applied in each experiment, and degradation parameters were monitored. Data analysis was performed using the "degradation tests analysis" technique, which is based on the concept of soft failure, i.e. the reaching of a threshold value of one of the monitored parameters. This approach assumes that a relationship exists between the value of the monitored parameter and the remaining useful life. The results of the study show the possibility of obtaining reliability estimates even in the absence of hard failures. These monitored parameters, in such cases, are related to failure modes that would not emerge in the usual product qualification testing or in Accelerated Life Testing.

1. Introduction
Every company aiming at customer satisfaction, tries to offer a functional product, safe and reliable. The needs of the customer concern the quality of the product, its cost and reliability. The increasing competition in the global market, besides , is forcing companies to reduce the time-to-market (TTM) of the product.
New products must be introduced into the market in the shortest possible time, with innovative features and a competitive cost[1], without compromising their efficiency and reliability[2].
Well, it is increasingly difficult to assess the duration of reliable products through the traditional lifetime testing, which monitors only the time-to-failure. In fact, in these cases, the traditional analysis cannot, in relatively short times,

[1] William Q. Meeker and Ying Shi, "Planning Accelerated Destructive Degradation Test with Competing Risks," 2010.
[2] Kulwant S. Pawar, Unny Menon, and Johann C. K. H. Riedel, "Time to Market," *Integrated Manufacturing Systems* 5, no. 1 (March 1, 1994): 14–22, doi:10.1108/09576069410815765.

either identify the product weaknesses (due to design or production), or confirm the predictions regarding the useful life[3].
To get information in a short time, many manufacturers have joined to conventional reliability testing Accelerated Life Test (ALT) and Accelerated Degradation Test (ADT).
ALT is a test in which the intensity of the applied stress exceeds normal use, whit the purpose of reducing the time needed to observe the effect of stress on the item. Through these tests, the mechanisms that determine the fault are accelerated and the time before the breakdown is reduced.
In ALT, we study hard failures, i.e. failures that have affected permanently the use of the product. If at the end of an ALT the number of hard failures is low, it is difficult or impossible to analyze life data and to make meaningful inferences on the reliability of the product. Thus, you can arrange, right from the beginning, an ADT which allows to analyze soft failures. These are the damages caused by the degradation that will eventually lead to failure and malfunction.
The common purpose of these two kinds of tests is the determination of the failure curve. More generally, starting from the test data, they try to deduce the device reliability under normal conditions of use.
The literature shows a significant development of the ADT in recent years. It should be noted, also, that there is no evidence of ADT applied to washing machines, one of the most common household appliances in the world. Even applications of ALTs in this area are rare[4].
Normally, in fact, these analyzes are dedicated to simple components and not to complex mechanical systems.
This article seeks to fill some of the gaps in the literature by extending ADT to the field of households appliances.
In this study, accelerated degradation tests have been carried out on a new washing machine, with an important technological innovation , in order to get estimates of the reliability parameters, focusing especially on the warranty period.
The use of such methodology for washing machines is original and brought interesting results, in terms of reliability performance acquaintance. Last but not least, a deeper expertise of the new product was quickly achieved.

2. ADT in reliability evaluation

Only in recent years ADT was applied to engineering fields[5] to evaluate and assess product reliability.
Previously, in the 80s, ADT was used in other sectors, for example, in biology. These tests, for instance, have been used to estimate the speed of biological degradation[6].

[3] Chen-Mao Liao and Sheng-Tsaing Tseng, "Optimal Design for Step-stress Accelerated Degradation Tests," *IEEE Transactions on Reliability* 55, no. 1 (March 2006): 59 – 66, doi:10.1109/TR.2005.863811.
[4] Sang-Jun Park et al., "Reliability Evaluation for the Pump Assembly Using an Accelerated Test," *International Journal of Pressure Vessels and Piping* 83, no. 4 (April 2006): 283–286, doi:10.1016/j.ijpvp.2006.02.014.
[5] W. Chen et al., "Accelerated Degradation Reliability Modeling and Test Data Statistical Analysis of Aerospace Electrical Connector," *Chinese Journal of Mechanical Engineering* 24, no. 6 (2011): 957.

Later, accelerated degradation tests, were applied in the industrial field, especially in electronics and electricity.
Degradation analysis methods are widely applied also to LEDs (light emitting diodes). These devices are relatively young and can be operated for a long time, even tens of thousands of hours consecutively. Therefore, they require special tests to verify their reliability without waiting years: several authors report examples of tests on these devices.
In mechanical engineering the ADT studies present in the literature are not as numerous as in electrical engineering. In a recent article the authorshows , through the use of data degradation, the reliability assessment of a train wheel using a simple linear model of the performance decay[7].
The influence of more than one performance on the failure mechanism is the basis of a recent study[8]. A multivariate analysis model is proposed, which entangles the observation and analysis of the statistical results of a variable at a time. In carrying out degradation tests on LEDs, a dependency between the several system performance is supposed. This leads to a better reliability estimation compared to estimates in which each performance is independent from the others.

3. Methods

ADTs evaluate the performance and the degradation properties of a component or of a product, caused by high stress. For each feature of the piece being tested, it should be possible to define the degradation suffered and, after stating the threshold beyond which the failure occurs, the time to failure of the unit can be inferred. This is defined, in fact, as the time in which performance degrades below a specified level or in which degradation reaches the threshold.
Degradation can be a physical parameter of the product, such as the corrosion of a plate of metal, or more simply, may be a performance, such as the brightness of a LED.
Nelson[9] first observed that the performance of many products worsen as their lifetime. This degradation, generally slow, can be accelerated by a stress higher than the user level one. In his article, he showed how the strength to break an electrical connection depends on both time and temperature.
The acceleration of stress can have various forms, as shown in Figure 1.

---

[6] T.B.L. Kirkwood and M.S. Tydeman, "Design and Analysis of Accelerated Degradation Tests for the Stability of Biological Standards II. A Flexible Computer Program for Data Analysis," *Journal of Biological Standardization* 12, no. 2 (April 1984): 207–214, doi:10.1016/S0092-1157(84)80055-4.
[7] Marta A. Freitas et al., "Using Degradation Data to Assess Reliability: a Case Study on Train Wheel Degradation," *Quality and Reliability Engineering International* 25, no. 5 (2009): 607–629, doi:10.1002/qre.995.
[8] Sari et al., "Bivariate Constant Stress Degradation Model: LED Lighting System Reliability Estimation with Two-stage Modelling," *Quality and Reliability Engineering International* 25, no. 8 (2009): 1067–1084, doi:10.1002/qre.1022.
[9] Wayne Nelson, "Analysis of Performance-Degradation Data from Accelerated Tests," *IEEE Transactions on Reliability* R-30, no. 2 (n.d.): 149 –155, doi:10.1109/TR.1981.5221010.

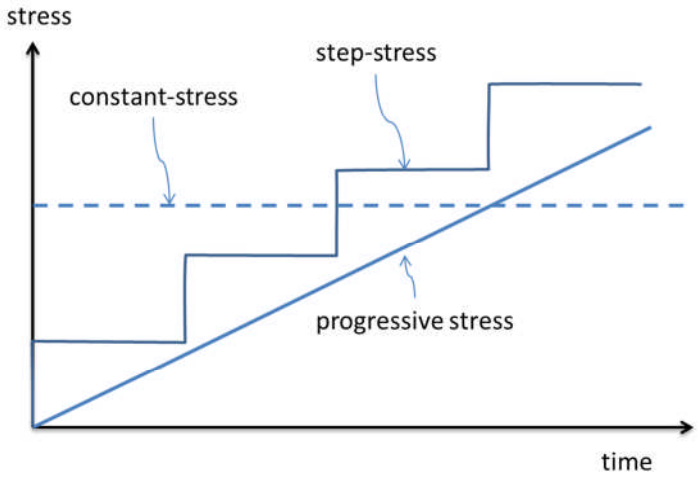


Figure 1: Various forms of stress acceleration in *ADT*. Depending on the type of stress acceleration, we can define three different test degradation types: a)constant-stress; b)step-stress; c)progressive-stress.

In the ADT with constant stress over time, the samples are tested under conditions of constant stress, for the entire duration of the experiment. A variation of this method was applied in our research.
Let's assume that the ADT is executed on n units and that during the test each of these is controlled periodically to measure the critical feature, y_{ij},where i indicates the unit considered and j is the time at which the inspection is performed. The equation of the degradation is (1):

$$y_{ij} = g(t_{ij};\ \beta_{1i};\ \beta_{2i};\ \dots;\ \beta_{pi}) + e_{ij} \qquad (1)$$

Where $g(t_{ij};\ \beta_{1i};\ \beta_{2i};\ \dots;\ \beta_{pi})$ is the true function of degradation of unit i at time t_{ij}.
$\beta_{1i};\ \beta_{2i};\ \dots;\ \beta_{pi}$ are the unknown parameters of the model, and e_{ij} is the term that indicates the error.
The units tested can undergo degradation and cross the threshold in various ways and at different times. In general, however, as time goes by, the degradation rises and therefore the probability that the values of the critical feature would exceed the default threshold[10].
If we call G the threshold value and T the time in which the fault occurs, the probability that a failure occurs before a certain time t is defined in Equation 2:

$$F(T) = P(T \leq t) = P[y(T) \geq G] = P[g(t_{ij};\ \beta_{1i};\ \beta_{2i};\ \dots;\ \beta_{pi}) \geq G] \qquad (2)$$

The likelihood of a fault depends on the parameter performance which, in turn, is a function of the stress level. The degradation tests are often performed by applying very high levels of stress with the aim of further accelerating failure mechanisms, and generating a greater number of damaging events. This makes it possible to reduce the uncertainty of the test results.

[10] Guangbin Yang, *Life Cycle Reliability Engineering* (John Wiley & Sons, 2007).

To perform ADT is desirable to define in advance the maximum threshold value above which faults may arise. The threshold may result from experience, design, or by experimental measurements. In any case, once made the tests, it is possible to fine-tune it.
In the present study, the ADT was performed with the aid of commercial software. With *ReliaSoft ALTA6*, in particular, the time to failures were extrapolated from the degradation parameter trends.
The next step is to measure the degradation over time. With time based regressions, you can extrapolate the degradation measurements until the predicted product failure.
In order to extrapolate the degradation measurements and to estimate the time-to-failure, you can adopt a linear, or an exponential, or a power law or a logarithmic base model (see Table 1).

Model	Expression
Linear	$l = a \cdot s + b$
Exponential	$l = b \cdot e^{as}$
Power law	$l = b \cdot s^{a}$
Logarithmic	$l = a \cdot \ln(s) + b$

Table 1. extrapolation models for the time to failure. l represents the nominal life, s is the stress and a and b are parameters of the model

Once calculated a and b, if we call s_i a generic stress level (with $1 < i < j$), we can extrapolate a MTTF $l(s_i)$. This, together with the other $l(s_j)$ can be used as a point in a data analysis to get the useful life.
The result of an ADT analysis, is summarized in a graph that describes the degradation trend of the critical feature and indicates the failure times deduced.

4. Case study

Having the need to evaluate, in shot times, the reliability performance of a new washing machine, we decided to assess if the DTA approach could be effective, compared to traditional approaches.
We chose a constant stress model for out ADT. The performance was verified by alternating overstress test with test cycles in which the stress was equal to the maximum value reachable by the user. This approach avoided the risk of measuring elastic over-deformations due to overstress, while our focus was on measuring the permanent deformation of the oscillating unit.
Going in particulars, ADTs have been designed and executed on the oscillating assembly of a new model of washing machine, with the purpose of assessing its reliability during the warranty period, namely the first two years.
The aspect that distinguishes the new washing machine is the higher load capability in the same space. This poses problems of slenderness of the mechanical parts and requires innovative materials.
Previous reliability studies showed that the critical mechanical components are inside the oscillating group. In addition, they have historically required

more technical service than other components. Since the new model is an evolution of the previous one, considering the increased dimensions of the oscillating group, it seems natural that this would still be the main responsible for the worsening of the machine.
In particular, in order to assess the reliability during the warranty period, we decided to test the washing machines for 500 cycles of 30 minutes each, corresponding to about two years of washes.
Since the vast majority of failures historically detected was mechanical, it was decided to overstress the new oscillating group by the imbalance increase of the load in the drum. With this stress, the deformation of the drum is amplified and it gets closer to the tub.
To optimize the duration of the tests, the cycle was reduced to the only spinning step, wherein the biasing mechanism is more present. All the phases with no influence on the searched failure were cut off. The duration of each washing cycle was so reduced from 1.5 hours to 30 minutes. The wash program selected is the most widely chosen by consumers: 60°C cotton wash[11]. The sample consisted of 24 washing machines, divided into groups of 8, subjected to three stress levels, all above the maximum imbalance of the drum permitted in normal use (400 grams). The three levels of stress applied in the three runs ware: 650 g, 800 g and 950 g (see Figure 2).
The study mainly analyzed two variables related to the degradation of the item tested. We have chosen the following measures:
• the distance between tub and drum;
• the outside temperature of the bearing positioned on the hub.
The measuring points chosen were: 3 for the distance drum-tub (outer circumference, median and internal), 2 for the temperatures (inner and outer bearing).
For the execution of ADTs and of the following analysis, it was necessary to identify the threshold values of the monitored variables. With a finite element analysis, it has been determined that the threshold value for the distance drum-tub is 3 mm. For the bearing temperature, a threshold value of 60 °C was set up taking into account the material properties and FEM model.
The test was time censoring, i.e. terminated for a fault or upon reaching the fivehundredth cycle.
The drum-tub distance wasn't measured in real time. Every 50 cycles, a measurement cycle was inserted, carried out with a normal unbalance of 400 g, in order to measure the distance. This approach allowed us to analyze the machine behavior (wear dependent) in normal use conditions, however after undergoing a process of "aging" due to the application of higher loads.

[11] O. Borgia, F. De Carlo, and M. Tucci, "Warranty Data Analysis for Service Demand Forecasting: A Case Study in Household Appliances," in *Advances in Safety, Reliability and Risk Management - Proceedings of the European Safety and Reliability Conference, ESREL 2011*, 2012, 1523–1529.

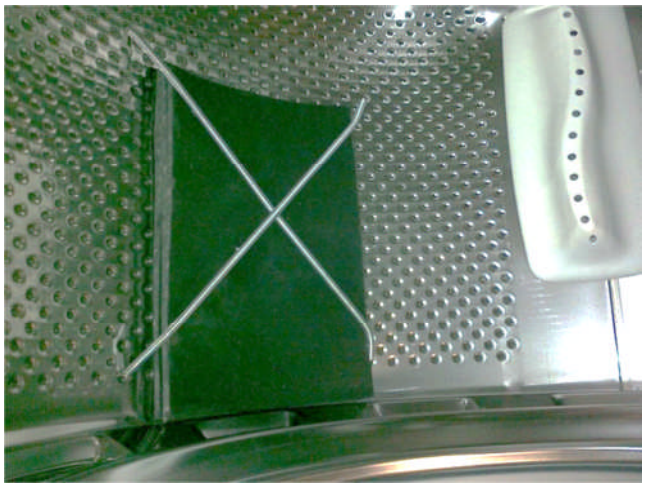


Figure 2: imbalance of the drum. The imbalance was induced by applying some pads of reinforced rubber, precisely positioned in the drum, to form a concentrated imbalance, causing the greatest stress.

To measure drum-tub distance, four distance sensors were used: three, staggered, placed on the upper part of the tank, and one positioned in the rear. The bearing temperature was continuously monitored, by averaging the values of two thermocouples.

5. Results

The drum, during the motion, undergoes a reversible deformation: from the initial cylindrical shape it tends to flatten, for the centrifugal forces of the motion. This phenomenon is emphasized as the unbalance increases. For this reason, the section of the drum becomes nearly an ellipse.
The distortion entity changes along the side wall of the tub. Therefore the distance drum-tub varies slightly depending on the axial section in which it is measured. For this reason it was decided to measure this distance in three different points.
For the calculation of the mean square error (MSE), we have chosen, as the reference distance, the average of the first 100 values of the distance. They were measured during the first operating cycle of the washing machine, when no deformation had yet began.
In Figure 3 degradation parameter patterns of the distance drum-tub are shown for one of the eight machines tested.
It can be noted that Ch3 values are the lowest. Ch1, positioned in the front part, registered the highest values. This is in agreement with the previous considerations: in the front of the oscillating group, the stresses are the greatest.

5.1 Drum-tub distance

When the deformation increased because of the stress, we could observe the increase in both the minimum and maximum distance and of the range. In Figure 4, MSE trends during the tests are displayed.
It’s observable that the MSE trend is rising with the number of cycles performed. That’s consistent with the physical phenomenon of degradation.

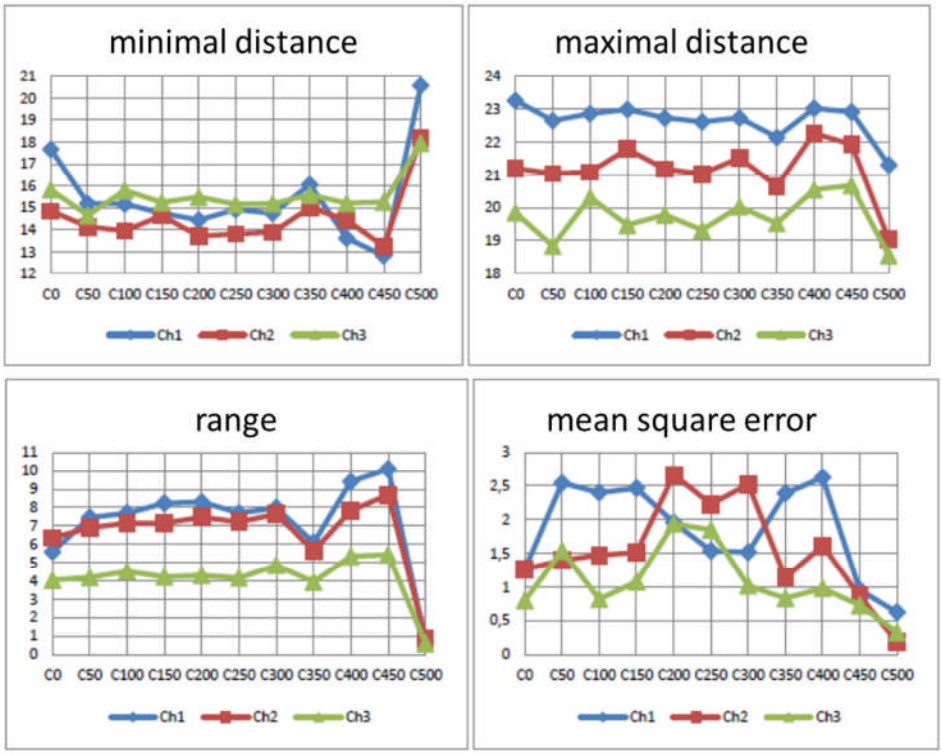


Figure 3: evolution of the degradation parameters for the drum-tub distance for one of the eight washing machines tested. Ch1, Ch2, Ch3 are the three different sensors, positioned respectively in the front, the rear and in the central part of the tub. The abscissa axes show the cycles.

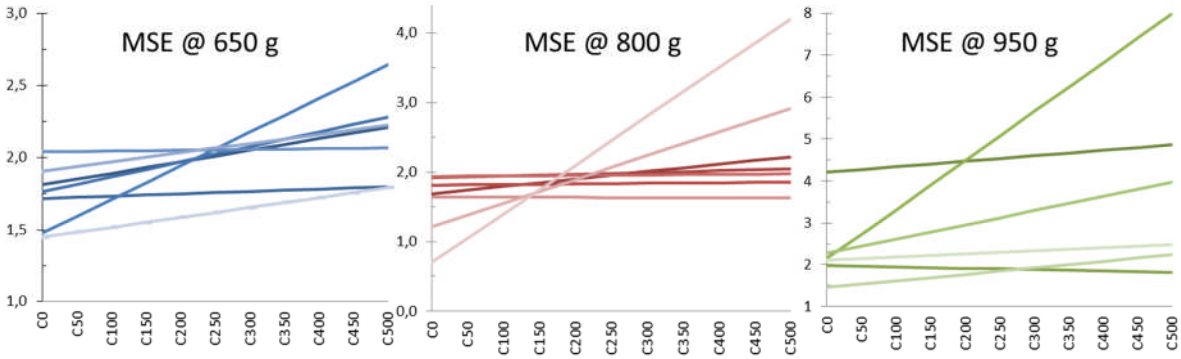


Figure 4: MSE of the drum-tub distance. In the diagram you can see the trends of the MSE, respectively for the three stress levels of 650g, 800 g and 950 g, for all the washing machines tested.

5.2 Bearing temperature

The temperature acquisition system allowed us to monitor and record values during all cycles performed by the washing machine, and not only every 50 cycles, as in the previous case. The time shape of temperature is oscillatory.
The degradation parameter associated with the temperature is the maximum value of an entire usage cycle, coincident with the centrifuge temperature during the measurement cycle. The analysis of the maximum temperature,

subsequently allowed us to evaluate the degradation suffered by the bearing while in motion, having taken 60 ° C as a threshold value (see Figure 5).

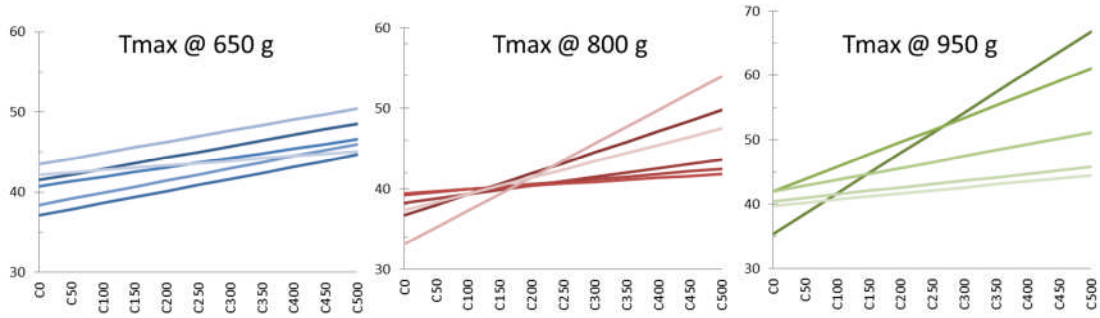


Figure 5: Maximum temperature of the bearing. In the picture you can see the trends of the highest bearing temperature, respectively for the three stress levels of 650g, 800 g and 950 g.

Even in this case, we can note the consistency between the maximum temperature and the physical degradation phenomenon: the maximum value is proportional to the number of cycles. The higher the slope of the straight line, the more accelerated is the mechanism of degradation.

6. Results analysis and Discussion

Aggregating the data of each stress level, we got the trend of the average values of the parameters of degradation over time.
So, a linear regression of the mean values was performed. Finally, to compare the results with each other, it was decided to normalize the values of each line relative to its initial value, leading to results seen in Figure 6.

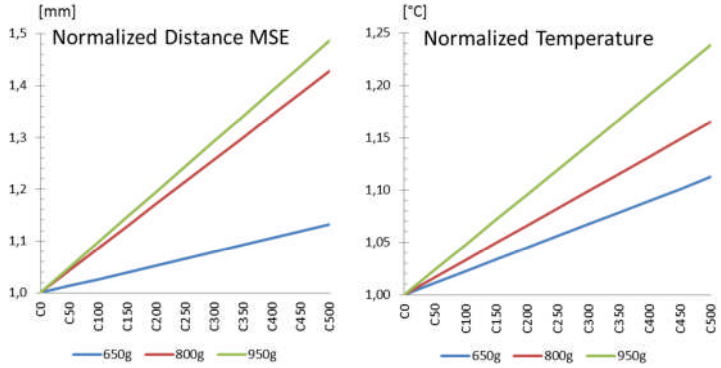


Figure 6: trend of normalized degradation parameters. The diagram shows the trends of: the MSE of the drum-tub distance and the maximum temperature. The values are normalized to their initial value, in order to compare them on a unique scale.

A first data observation leads to note that the slope of the straight lines is higher for washing machines with greater imbalance. This denotes a more rapid increase of the parameter and, consequently, a greater degradation process.

6.1 Drum-tub distance

The boundary value of the standard deviation was found during testing. In fact, during a test, there was a break in a washing machine and the value of the minimum distance just before the fault was equal to 3.2 mm. It is very interesting to note the remarkable correspondence between the theoretical value, based on FEM analysis and the experimental one.
All MSE test values were inserted in the software to perform the data analysis and to extract reliability information are visible in Figure 7.

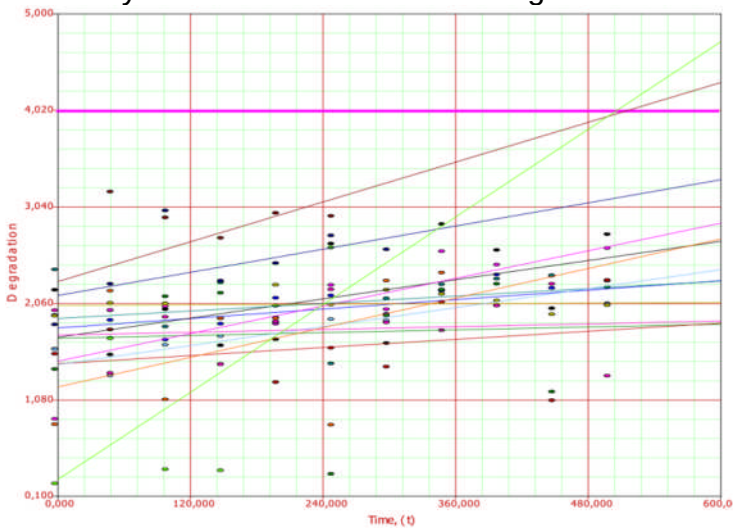


Figure 7: drum-tub distance MSE. The graph shows the relationship between degradation and time: on the abscissa there are the cycles and on the ordinate the degradation parameter values. The big horizontal line is the threshold of the standard deviation, 4 mm. The other lines show the parameter trend of each machine.

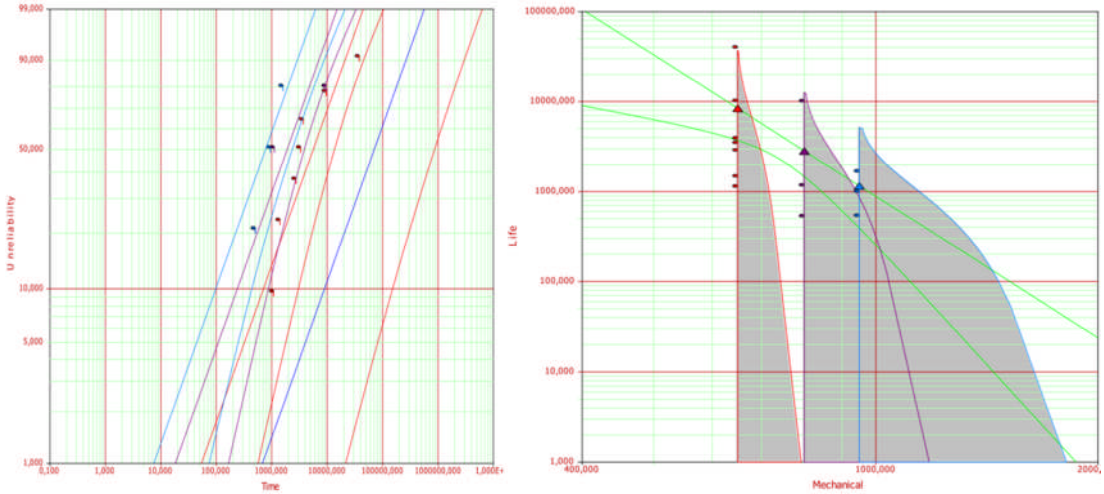


Figure 8: Reliability evaluation. On the left side you can see the soft failure points, in the three different conditions of accelerated stress, fitted by a Weibull distribution. In the right part, the same points are visible in the graph showing the relationship between life and stress.

It has been possible to identify, for the normal use with 400g of unbalance, a Weibull distribution with shape parameter β = 0.91 (Figure 8). This parameter indicates a typical decreasing failure rate device.

6.2 Bearing temperature

It was then studied the trend of the maximum temperature of the bearing. The degradation caused by the imbalance involves an increase in the temperature, which was measured in normal load conditions.

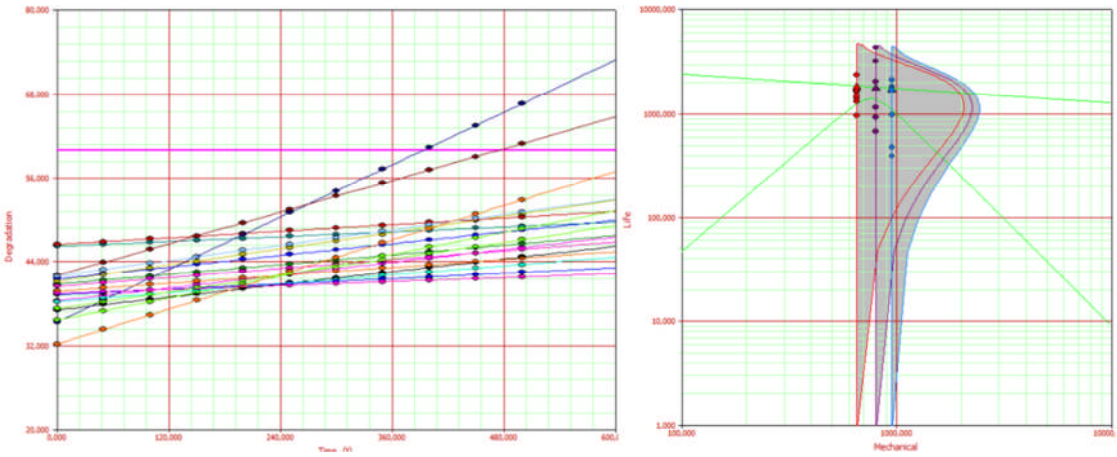


Figure 9: temperature soft failures. The diagram shows, on the left side, the lines interpolating the temperatures measured during the tests. From the intersection with the threshold value of 60 °C the soft failures were obtained. These latter are visible in the right diagram, grouped by stress levels.

The reliability evaluation in normal load conditions, led us to define a Weibull function with shape factor $\beta = 1.80$. This signifies an increase in the failure rate over time, due to aging. We could calculate the reliability value after 500 and 1250 cycles, which was worse than the ones calculated by the distance drum-tub analysis.

A drastic reduction of the MTTF was also observed in temperature based evaluations, if compared to the results gained from the distance drum-tub. A possible explanation is that the threshold value of 60°C, obtained with FEM analysis as well, may be not correct. In fact, during the tests, it has been reached many times but not any damage happened.

Conclusions

The purpose of this work was to verify the possibility of quickly obtaining reliability information on a new type of washing machines. The high intrinsic reliability of the product led us to apply the theory of ADTs, although it is not present in the literature any application to similar products. The analysis focused on a couple of degradation parameters: the distance of drum-tub and the bearing temperature.

An excellent result was that a break has occurred immediately after that the deformation parameter had reached the FEM threshold of the distance drum-tub. Not the same result was gained with the temperature, in which the threshold was has been exceeded a few times without consequences.

The methodology let us to evaluate the reliability function of the oscillating group. It should be noted that this kind of assessment would not have been possible with standard techniques for reliability estimation.

Finally, it should be emphasized that the execution of measuring cycles under conditions of maximal user load, allowed us to study the washing machines as artificially aged. This made it possible to identify their weak points, before

putting on the market. Finally the analysis allowed the company to redesign some new components which resulted weaker and to modify the quality acceptance testing in order to increase the reliability of the product.

References

1. Borgia, O., F. De Carlo, and M. Tucci. “Warranty Data Analysis for Service Demand Forecasting: A Case Study in Household Appliances.” In *Advances in Safety, Reliability and Risk Management - Proceedings of the European Safety and Reliability Conference, ESREL 2011*, 1523–1529, 2012.
2. Chen, W., J. Liu, L. Gao, J. Pan, and S. Zhou. “Accelerated Degradation Reliability Modeling and Test Data Statistical Analysis of Aerospace Electrical Connector.” *Chinese Journal of Mechanical Engineering* 24, no. 6 (2011): 957.
3. Freitas, Marta A., Maria Luíza G. de Toledo, Enrico A. Colosimo, and Magda C. Pires. “Using Degradation Data to Assess Reliability: a Case Study on Train Wheel Degradation.” *Quality and Reliability Engineering International* 25, no. 5 (2009): 607–629. doi:10.1002/qre.995.
4. Kirkwood, T.B.L., and M.S. Tydeman. “Design and Analysis of Accelerated Degradation Tests for the Stability of Biological Standards II. A Flexible Computer Program for Data Analysis.” *Journal of Biological Standardization* 12, no. 2 (April 1984): 207–214. doi:10.1016/S0092-1157(84)80055-4.
5. Liao, Chen-Mao, and Sheng-Tsaing Tseng. “Optimal Design for Step-stress Accelerated Degradation Tests.” *IEEE Transactions on Reliability* 55, no. 1 (March 2006): 59 – 66. doi:10.1109/TR.2005.863811.
6. Meeker, William Q., and Ying Shi. “Planning Accelerated Destructive Degradation Test with Competing Risks,” 2010.
7. Nelson, Wayne. “Analysis of Performance-Degradation Data from Accelerated Tests.” *IEEE Transactions on Reliability* R-30, no. 2 (n.d.): 149 –155. doi:10.1109/TR.1981.5221010.
8. Park, Sang-Jun, Sang-Deuk Park, Kwang-Suck Kim, and Ji-Hyun Cho. “Reliability Evaluation for the Pump Assembly Using an Accelerated Test.” *International Journal of Pressure Vessels and Piping* 83, no. 4 (April 2006): 283–286. doi:10.1016/j.ijpvp.2006.02.014.
9. Pawar, Kulwant S., Unny Menon, and Johann C. K. H. Riedel. “Time to Market.” *Integrated Manufacturing Systems* 5, no. 1 (March 1, 1994): 14–22. doi:10.1108/09576069410815765.
10. Sari, Newby, Brombacher, and Tang. “Bivariate Constant Stress Degradation Model: LED Lighting System Reliability Estimation with Two-stage Modelling.” *Quality and Reliability Engineering International* 25, no. 8 (2009): 1067–1084. doi:10.1002/qre.1022.
11. Yang, Guangbin. *Life Cycle Reliability Engineering*. John Wiley & Sons, 2007.

A Petri-Net Modelling Approach to Rail Track Geometry Maintenance and Inspection

Matthew Audley and John Andrews
Nottingham Transportation Engineering Centre, University of Nottingham

Abstract

Maintaining track on the UK railway system accounts for a large proportion of the total operating costs. During the year 2009/2010 Network Rail spent £464 million on the maintenance of track and an extra £698 million on track renewals. In order to plan the investment required to maintain the railway in a good condition, models are used to predict the effects of different asset management strategies.

This study presents a Petri-Net modelling approach to rail track asset management. The rail track geometry is measured periodically by means of a measurement train in order to establish the quality of the track geometry. When the track geometry drops below an acceptable condition, maintenance is performed to restore the condition. Such maintenance can be achieved using either tampers or stoneblowers. However studies of the degradation process have shown that such maintenance becomes less effective the more times it is carried out and there will be a point at which ballast replacement becomes the most cost effective solution. Using the Petri net model to predict the track condition over time, accounting for a specified asset management strategy, the Whole Life Cost has been minimised. This has been accomplished using a genetic algorithm to determine the optimum: geometry measurement interval and maintenance and renewal criteria. Such a predictive approach would be used to support the setting of the maintenance strategy.

Keywords: Petri net, Tamping, Track Degradation, Track Asset Management, maintenance decision making tool, Life Cycle Analysis.

1. Introduction

The UK rail industry remains a safe, reliable and competitive alternative to other forms of transport. Maintaining such an aging, increasingly utilised, railway network with limited financial resources is a major challenge. A track state modelling approach can be used to develop an effective track asset management process.

The model presented in this paper, is based on the approach developed by [1] and later refined by [2-4], and uses the track degradation distributions introduced by [5].

2. Petri Net Methodology

An extensive literature review revealed two suitable modelling techniques for a track degradation model: Markov and Petri Net. The Markov model was

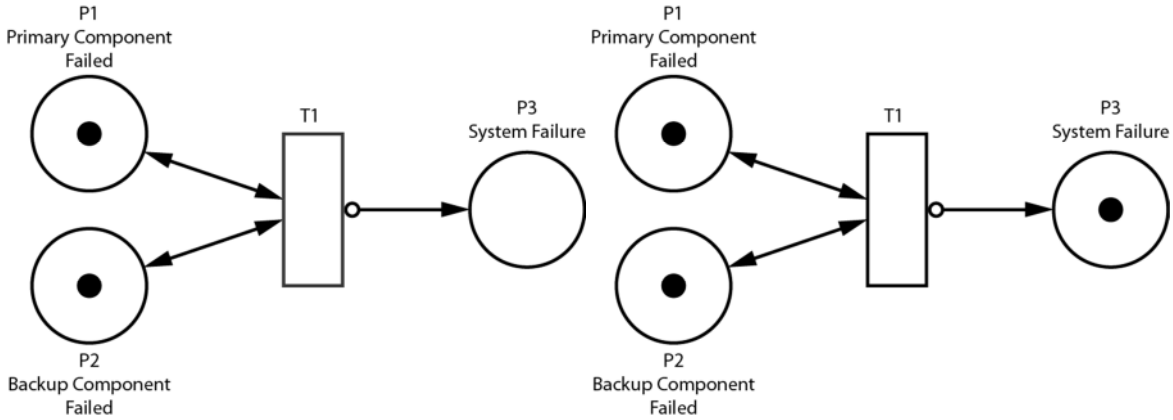
rejected on the basis that it requires a constant deterioration rate when transitioning from one state to another, whereas a Petri net model accepts any type of deterioration rate. This flexibility is an important characteristic for a track degradation model as rail track is known to deteriorate at a variable rate.

A Petri net (PN) is a directed bipartite graph where the nodes represent transitions and places linked by directed edges. They were first introduced by Carl Petri in 1962 and are used as a graphical representation of systems which contain one or more dynamic processes.

Petri Nets consist of four main components:

1. *Tokens* - May be created, destroyed or transferred from place to place via transitions. They are graphically represented by a filled in circle. However, tokens may be represented by a number if the number of tokens is so high that it is impractical to display them.
2. *Places* - Places are graphically represented by a circle and can store an infinite number of tokens. When occupied by a token(s) they reveal information about the current system state or condition.
3. *Transitions* - Move the tokens around the system, they appear as a rectangle on a Petri net model. A delay, distribution or probability of firing may be attributed to the transition. A transition may only fire if it is enabled.
4. *Arcs* - Link input places to transitions and transitions to output places using a directed edge to represent the movement of the token. Arcs have an associated multiplicity. By default, if no number is assigned then its multiplicity is one. Furthermore, an arc may contain an inhibitor edge (shown by a small circle on the transitions side) which prevents a transition from firing if the input place contains a token. Arcs are graphically shown by a line with at least one edge type.

Figure 1 shows a simple Petri net model of a system containing a primary component and a backup component, when both fail the system is said to have failed. In figure 1 (a) places P1 and P2 are marked indicating that the primary and backup components have failed. This enables transition T1 which subsequently fires, marking places P1, P2 and P3 as shown in figure 1 (b). Here P3 is marked indicating the system has failed. In order to prevent transition T1 from firing indefinitely an inhibitor arc connects P3 to T1.



(a) (b)

Figure 1. Basic Petri Net Model

2.1 Additional Transitions

A modified version of the standard Petri net model is used in this paper. The presented track model makes use of two additional transitions, a reset transition and a place conditional transition. The reset [1] transition operates in the same manner as a regular transition. However, in addition to the regular functions, upon firing, the reset transition will set certain places within the network to their original state. These additional functions of the reset transition can be performed using the standard Petri net features. However, their use reduces the size and complexity of the model. Reset transitions are accompanied by a list of the places that they reset. The place conditional transition performs the regular functions of standard transition. However, the distribution from which durations / probabilities are sampled is dependent on the number of tokens residing in one or more places within the Petri net. These places are connected to the place conditional transition via a dashed arc. Place conditional transitions cannot be replicated using standard Petri net features.

3. Track Geometry Degradation

A measurement train passes over the UK rail network at regular intervals, typically once per month for the main lines. This train records several track geometry measurements of which the measurement considered to be the most representative of track quality is the vertical alignment. This measurement is expressed as the standard deviation of the averaged vertical position of the rails, typically a 35m wavelength filter is applied.

The track quality degrades with respect to time and passing traffic. The track quality does not improve on its own and requires corrective maintenance when the condition falls below certain standards. In severe cases speed restrictions or line closures are imposed until corrective maintenance or renewal is performed. Following an intervention the track does not return to an as new condition, its initial condition and the rate at which it degrades are dependent on its maintenance history, line speed and the quality of the work carried out [5]. In [5] it was noted that two probability distributions may be used to express track degradation, a Lognormal distribution and a Weibull distribution. The first represents the probability of achieving a given track quality following an intervention. The second is used to express the rate at which track degrades following an intervention. The method used to obtain these probability distributions is outlined in [5].

4. Track Policy

The averaged vertical alignment of the two rails is expressed as a standard deviation for a 220 yard track section. These standard deviations are assigned quality bands which in turn are used to describe the current track state. These quality bands are dependent on line speed and are set by the rail operator.

Once the track degrades to an unacceptable level and this level has been identified, maintenance is requested.

The maintenance option considered in this model is tamping.

Following maintenance the track's condition reverts to some initial state. It will then degrade at a rate which is dependent on the line speed and maintenance history.

A ballast renewal is carried out once maintenance is no longer economically viable. The model allows for renewal to be carried out once maintenance becomes too frequent. Following renewal the maintenance history of the line is reset.

5. Track section model

The Petri net model outlined below is made up of six modules. Three of these modules may be modified depending on which asset management policy is to be modelled. This is carried out by using enablers. Enablers are places that can be marked / unmarked in order to enable particular modules or represent asset management strategies. Besides the enabler places, all of the marked places within the modules shown in figures 2 to 6 are the default markings and must be marked prior to running any simulations. The enablers may be marked / unmarked depending on which asset management strategy is to be modelled.

5.1 Track Degradation

Figure 2 shows the track degradation module. Following maintenance/renewal the track is in some initial state as marked by place P1. Place conditional transitions T1 to T5 are then enabled and a probability of firing is sampled from a lognormal distribution. Depending on which transition fires the track state will range from very good to very poor as represented by places P2 to P6. The time taken to degrade to the next state is then sampled from a Weibull distribution.

The distribution from which probabilities or durations are sampled for transitions T1 to T4 and T6 to T9 is dependent on the maintenance history of the track described by place P26 and the line speed being modelled.

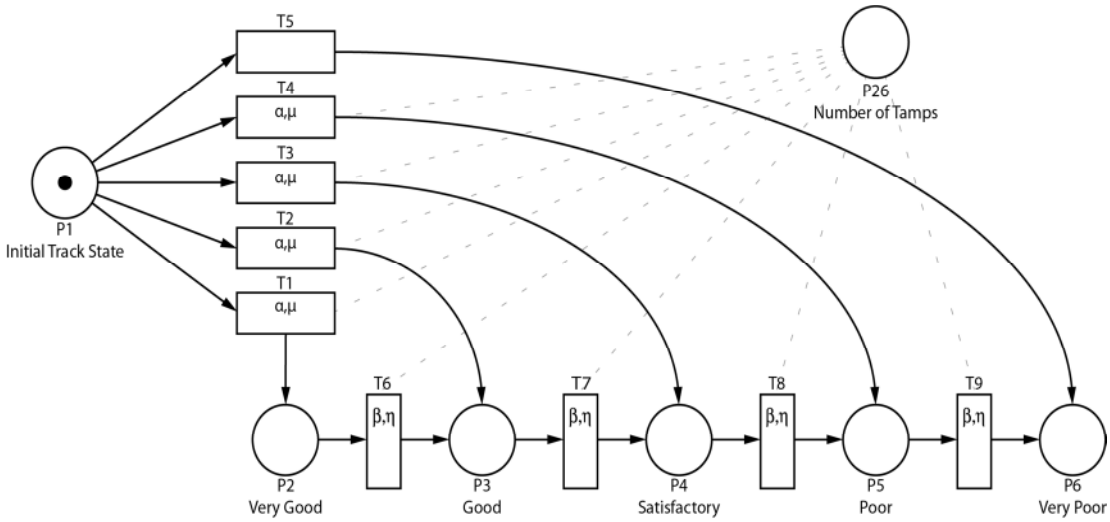


Figure 2. Track Section Degradation Module

5.2 Track Inspection

The track car inspection module [5] is shown in figure 3. The measurement train inspection cycle is represented by the loop P12 → T16 → P13 → T15. When P12 is marked T16 is enabled and an inspection will occur in θ days. After T16 fires, P13 is marked and inspection of the track section begins, the ε delay associated with T15 ensures that the track car cannot leave until the section has been inspected. If the track has degraded to P2-P6 since the last inspection, the relevant transition T10 to T14 will fire, marking P7 to P11. If the known state has changed since the last inspection two places will be marked. The relevant transition T17 to T26 will fire to remove the previously identified state. For example, If the outcome of the first inspection resulted in place P8 being marked and the second inspection reassessed the condition as being P11 then transition T23 will be enabled and upon firing T23 will remove tokens from P8 and P11 and place one token in P11. The condition of the track is now known.

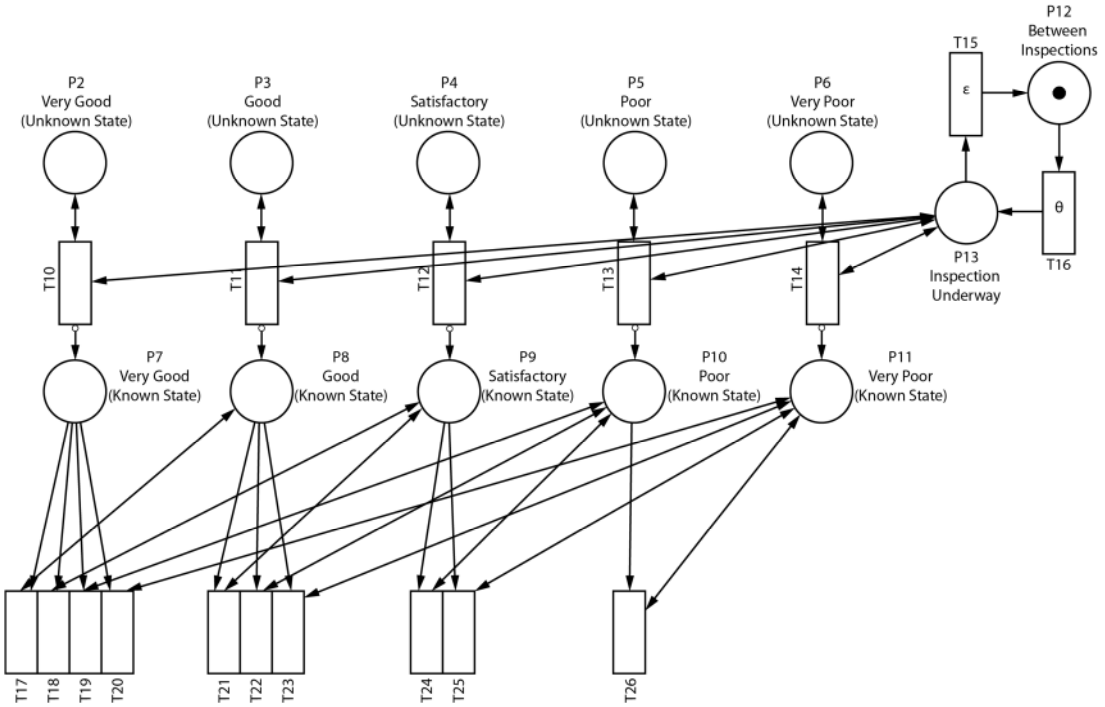


Figure 3. Track Inspection Module

5.3 Maintenance Policy

Figure 4 shows the maintenance policy module for the level of degradation at which maintenance is requested. This can be adjusted depending on which enablers P19 to P23 are marked. For example, in figure 4 places P21 to P23 are marked. Therefore, maintenance will only be requested when the track section degrades beyond a satisfactory state i.e a place between P9 and P11 is marked. Once this criteria is met the relevant transitions will fire and P25 will be marked.

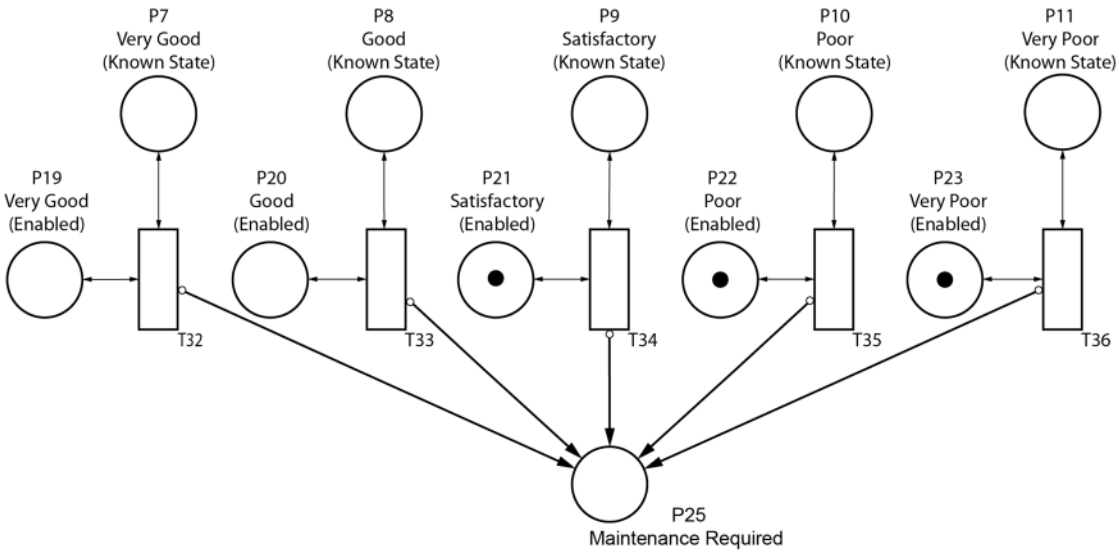


Figure 4. Interim Maintenance Policy Module

6. Interim Maintenance

Figure 5 shows the maintenance module. When P25 is marked maintenance is required on the track section and T37 is enabled. A time is sampled from the distribution of times taken to get a tamper to the track section. Once the time has elapsed T37 will fire and P27 will be marked. A token will also be added to P26 to indicate that a tamp has been carried out on this track section. P25 will remain marked as the maintenance work has not been completed.

Once P27 is marked T39 is enabled, after time t_{pr} it will fire marking P1. T39 is a reset transition, upon firing places P2-P11, P29 and P30 are reset to their default state.

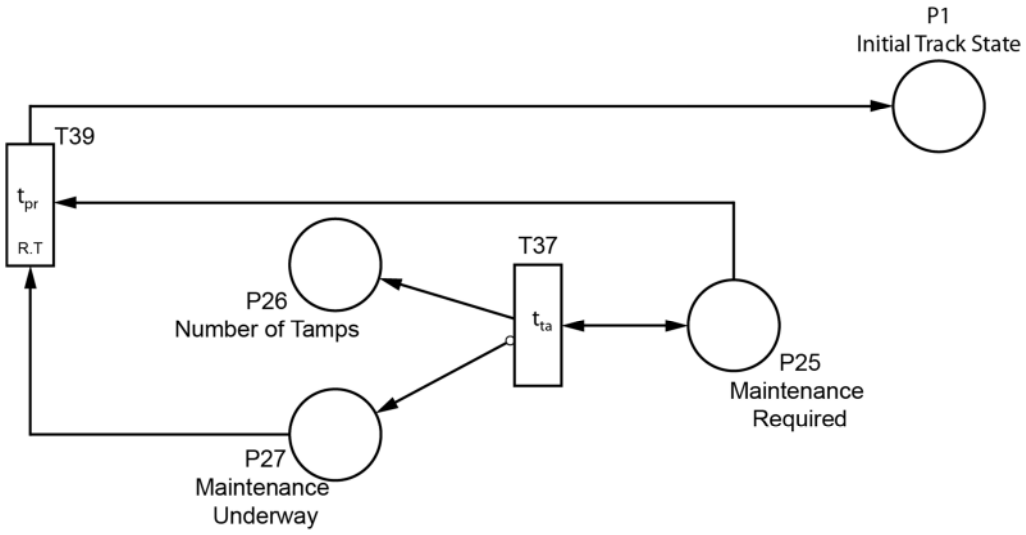


Figure 5. Interim Maintenance Module

6.1 Renewal Policy

A renewal is different to interim maintenance. A full renewal is replacement of the ballast, sleepers and rails.

Track is renewed once interim maintenance becomes too frequent. However, there is little understanding as to what is too frequent. Moreover, any interim maintenance required shortly after a prior maintenance action may just be down too sub-standard maintenance rather than poor track quality. The module shown in figure 6 is used to model the renewal.

If P31 is marked then the frequency based maintenance policy is enabled. First t_{bm} and n1 must be set. Once set, any maintenance which occurs within t_{bm} of a previous action is considered to be too frequent. This will incur one strike. Once the number of strikes is greater than n1, renewal will be performed.

If P25 is marked before T41 fires T45 will be enabled. Upon firing a token will be added to P34. If the number of tokens in P34 is greater than n1 T46 will be enabled and subsequently will fire. This removes all of the tokens from P34 and places a token in P35. This enables T47 and inhibits T39 in case any

maintenance is scheduled during this process. T47, which is a reset transition, will reset places P2 to P11, P26 and P29 upon firing.

If T41 fires the track section is no longer eligible for a renewal and P30 will be marked inhibiting T45. If any tokens reside in P34 then T44 will be enabled. T44 will fire until all of the tokens residing in P34 are depleted. T42 will fire after ε marking P33. Tamping may then be performed on the track once the associated places from figure 5 are marked. If no token resides in P31 the module is not enabled, transition T43 will be enabled marking P33 so that regular maintenance can be carried once T39's associated places are marked.

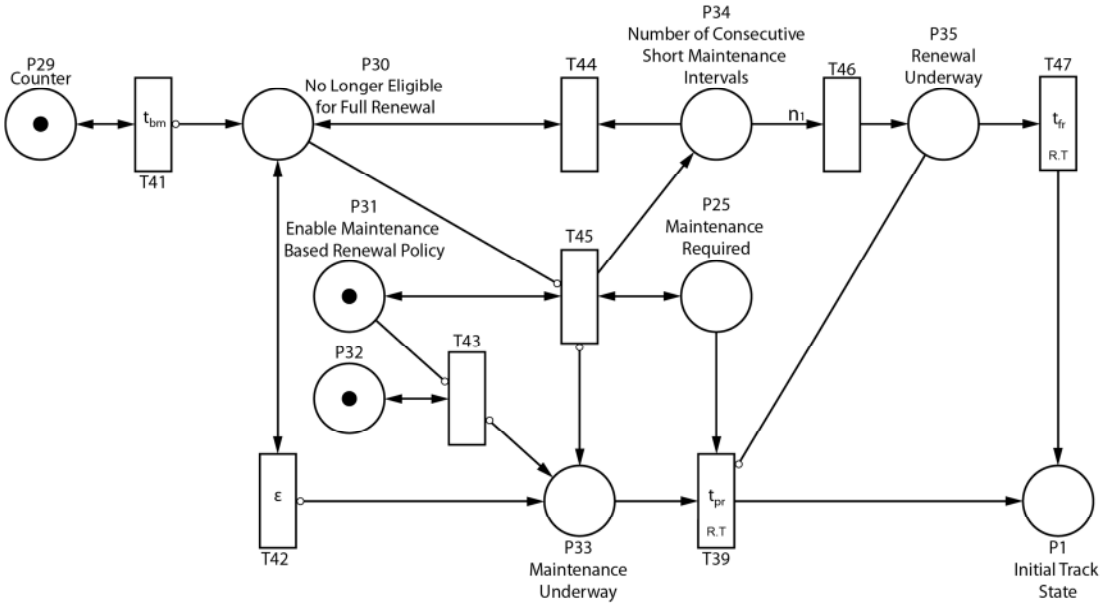


Figure 6. Frequency of Maintenance Based Renewal Module

7. Monte Carlo Sampling

Sampling from distributions is required for all of the stochastic transitions.

For the transitions using the lognormal distributions, the probability of firing is calculated as:

$$F(\sigma|\mu,\alpha) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^{\sigma} \frac{e^{\frac{-(\ln(t)-\mu)^2}{2\alpha^2}}}{t}\,dt \qquad (1)$$

A random number (X) between 0 and 1 is generated. If $X \geq f(\sigma)$ then the transition will fire. The order of firing starts with transition T1 and continues to T5. For example, if T1 does not fire then the same test is applied to T2 then T3 etc. If T1 to T4 do not fire then transition T5 will fire as there is no associated probability for T5.

For the transitions which feature a Weibull distribution, a random number X is also generated and equated to the cumulative probability:

$$F(t) = 1 - exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] = X \qquad (2)$$

This is rearranged to give:

$$t = \eta\left[-\left(ln(X)\right)\right]^{\frac{1}{\beta}} \qquad (3)$$

8. Simulation

Bespoke software was designed and written to solve this Petri net model. The model is executed until the simulation end time is reached. The simulation end time for the model was set at 50 years. The simulation is ran 1000 times.

Upon completion of the Monte Carlo simulation statistics are collected in order to establish the performance of the asset management strategy. The statistics collected are:

i. The number of tamps.
ii. The number of renewals.
iii. The time between renewals.
iv. The condition of the track at any time.

8.1 Costing

Costs are applied to the model in order to compare the asset management strategies.

If the track deteriorates to a poor condition, a speed restriction is imposed. If the track is allowed to degrade to the very poor condition, then a line closure will be applied to the track section. The rail operator will incur a fine if either of these is allowed to happen. The amount of the fine is based on the duration that the line spends in each of these states. Table 1 shows the costing information used for this analysis.

	Cost (units)
Measurement train	1
Tamping	100
Renewal	10000
Line Closure (per day)	40
Speed Restriction (per day)	20

Table 1. Table of Costs

9. Optimisation

The Petri net has been designed so that a number of asset management strategies may be trialled in order to find the most cost effective solution. There are a number of parameters within the Petri net which may be adjusted in order to represent a particular asset management strategy.

The track inspection car monitors the track condition at a certain time interval θ_1. If θ_1 is too large the length of time the track resides in an unknown condition increases. Hence, track may degrade to a condition requiring a speed restriction or line closure and the infrastructure operator would not be

notified. Conversely, the smaller θ_1 gets the time the track spends in a unknown condition decreases. However, there are a limited number of inspection cars on the network and buying extra measurement trains is expensive. Thus, if θ_1 is too small the track will not degrade very much between inspections. Hence, money will have been spent on extra measurement trains when not needed. Therefore, an optimum inspection time θ is sought. There are an infinite number of possibilities for θ; however, any solution must be sensible and achievable.

Figure 4 shows how the level of degradation at which maintenance is requested can be adjusted. Maintenance is costly, thus if maintenance is performed too soon money will not have been spent in a cost effective manner. Also maintenance damages the track ballast. Therefore, the track will degrade quicker following maintenance. There are five states for the maintenance option.

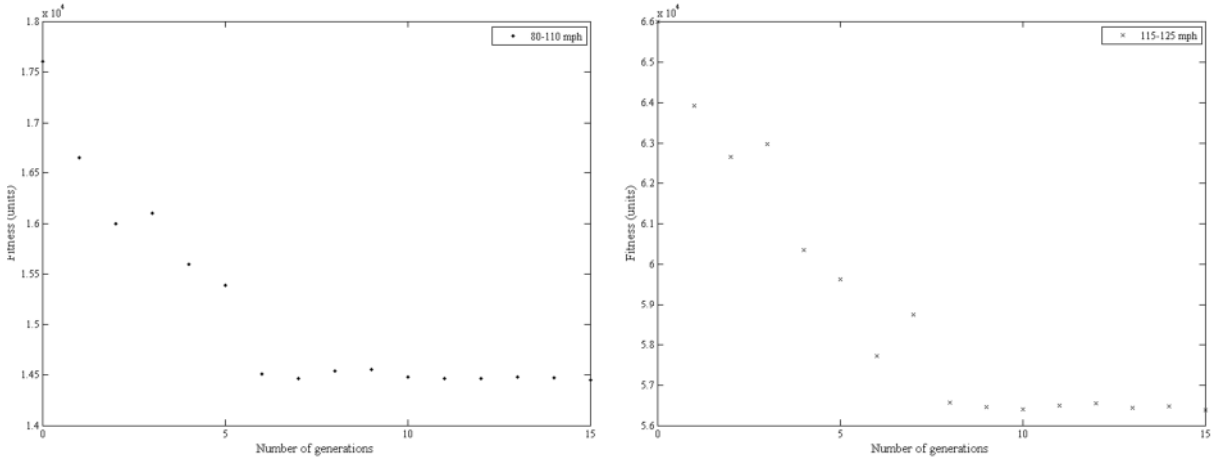
For the frequency based renewal policy shown in figure 6 there are two parts which may be optimised, the period of time, which is considered to be too frequent for maintenance (t_{bm}), and the number of consecutive too frequent maintenance actions (n_1). Once interim maintenance becomes too frequent it is no longer a cost effective option and a ballast renewal should be performed. What is too frequent is currently unclear. Furthermore, if maintenance occurs too frequently on only one occasion, then it may be down to poor maintenance as opposed to poor track. In this case a renewal would not be the most cost effective option. Therefore, the optimum number of consecutive too frequent maintenance actions (n_1) before a renewal, becomes cost effective is sought.

Table 2 shows the asset management strategy options trialled. There are 625 possible combinations for the asset management strategy. When accounting for the number of simulations required to gain convergence using the Monte-Carlo approach, an exhaustive investigation of these options is not practical. Therefore, an optimisation technique has been used. The genetic algorithm was found to be the most suitable.

Genetic algorithms are a form of evolutionary computing. A simulation is run with a particular asset management strategy (represented by chromosomes) called the population. Solutions from one population are taken and used to form a new population. Solutions which are selected to form new solutions (offspring) are selected according to how suitable they are (fitness). The more suitable they are the more chances they have to reproduce. This is repeated until some criteria are satisfied, in the case of this study this will be the lowest mean life cycle cost. The optimum asset management policy is found for two speed bands, 80-110 mph and 115-125 mph. This was carried out using a population size of 15, a mutation rate of 0.01, a crossover fraction of 0.8 and an elite count of 2. Where elite count is the number of individuals with the best fitness values in the current generation that are guaranteed to survive to the next generation and crossover fraction is the percentage of the population (minus elites) which will become crossover children. For our case there will be 10 crossover children at each generation. The genetic algorithm is stopped after 15 generations. The Petri net simulation was ran 1500 times for each iteration.

		Maintenance Frequency Based Renewal	
Inspection interval (Days)	Maintenance Level	Time	Frequency
7	Very Good	240	1
14	Good	360	2
28	Satisfactory	480	3
35	Poor	600	4
56	Very Poor	720	5

Table 2. Asset Management Policy Options



(a) 80-110 mph (b) 115-125 mph

Figure 7. Genetic Algorithm results after 10 generations

				Maintenance Frequency Based Renewal	
		Inspection interval (Days)	Maintenance Requested	Time	Frequency
Speed Band (mph)	80-110	14	Satisfactory	480	4
	115-125	14	Poor	360	3

Table 3. Genetic Algorithm Results

10. Results

The following results demonstrate the capabilities of the model. Figures 7a, 7b and Table 3 show the outcome of the genetic algorithm. Figure 7a shows that a optimal solution for the 80-110 mph line speed was obtained after 6 generations and an optimal solution for the 115-125 mph line speed was obtained after 8 as shown in figure 7b. The fitness value fluctuates even though the solution remains the same due to the fitness value being based on

a mean cost. For both speed bands, the inspection interval (14 days) is equal. 1500 simulations for each iteration was more than enough as the results started to converge much earlier, as shown in figures 8 and 11. Figure 11 shows the mean cost for the two different asset management policies for the two different line speeds. The cost of the 115-125 mph speed band is around 3 times more expensive than the 80-110 mph speed band.

80-110 mph speed band

For this line speed, the optimum renewal policy found by the genetic algorithm was to perform a renewal if maintenance occurs within 480days of a previous maintenance action on 4 consecutive occasions. The mean number of tamps converges to approximately 4 and the mean number of renewals to 1 after 50 years as shown in figures 8a and 8b respectively. This results in a mean of 4 tamps per renewal and a mean time of 40 years between renewals, as shown in figures 8c and 8d respectively.

Figure 8a shows the condition distribution plot for the 50 year period.

At year 0 the track has just been renewed. However, renewing track does not guarantee it being in the very good condition state. Furthermore, there is a slight chance that the track may be in a very poor condition following renewal. In addition, the probability of being in a very good condition decreases with time whilst the probability of being in a good to very poor state increases with time.

115-125 mph speed band

For this line speed, the optimum renewal policy found by the genetic algorithm was to perform a renewal if maintenance occurs within 360 days of a previous maintenance action on 3 consecutive occasions. The mean number of tamps performed on the track within a 50 year period is approximately 14 and the mean number of renewals is 3.4, as shown in figures 8a and 8b respectively. This equates to a mean of 4.3 tamps per renewal and a mean time of 15 years between renewals, as shown in figures 8c and 8d.

Figure 9b shows the condition distribution plot for the 50 year period. Again, at year 0 the track has just been renewed, and following renewal, there is no guarantee as to what condition the track will be in. The probability of being in a very good condition decreases with time whilst the probability of being in a good, satisfactory, poor or very poor state remains approximately constant.

Comparison

Figure 10 shows the mean percentage of time spent in each state. The track section model for the 80-110 mph speed band spent 80% of its time in the "Very Good" state where as for the 115-125 mph speed band the time spent in the first three condition states is more evenly spread. This implies that it is easier to retain a better track quality for the lower line speed.

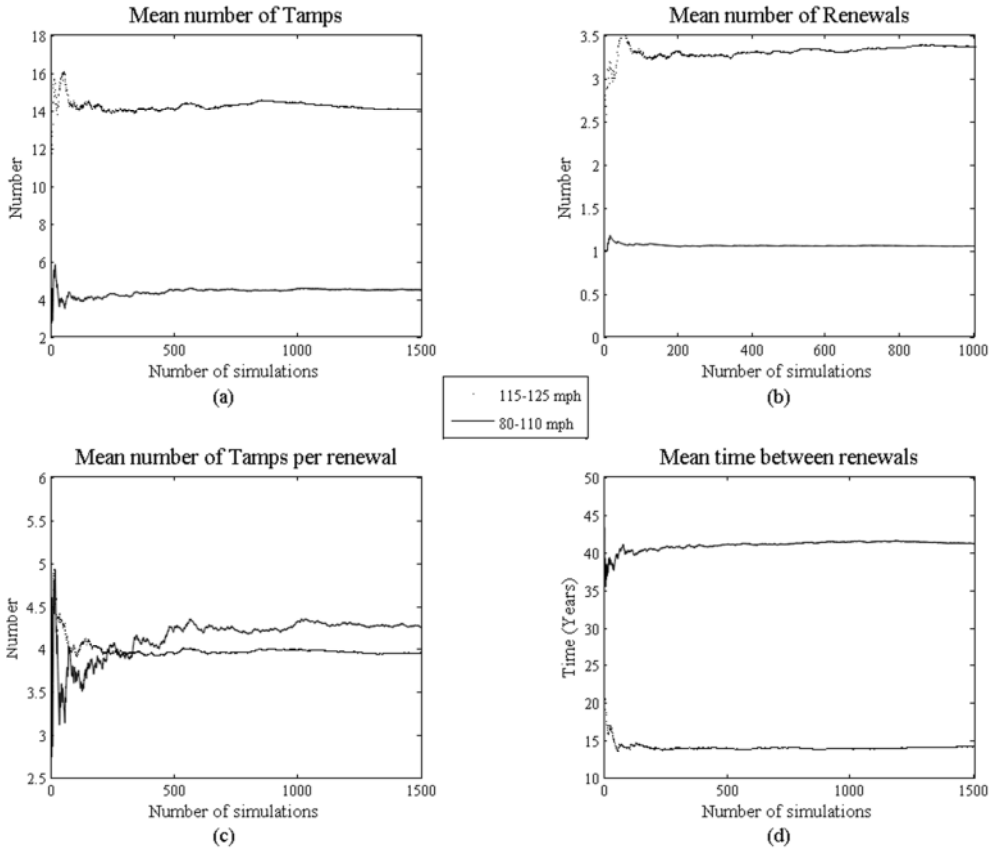
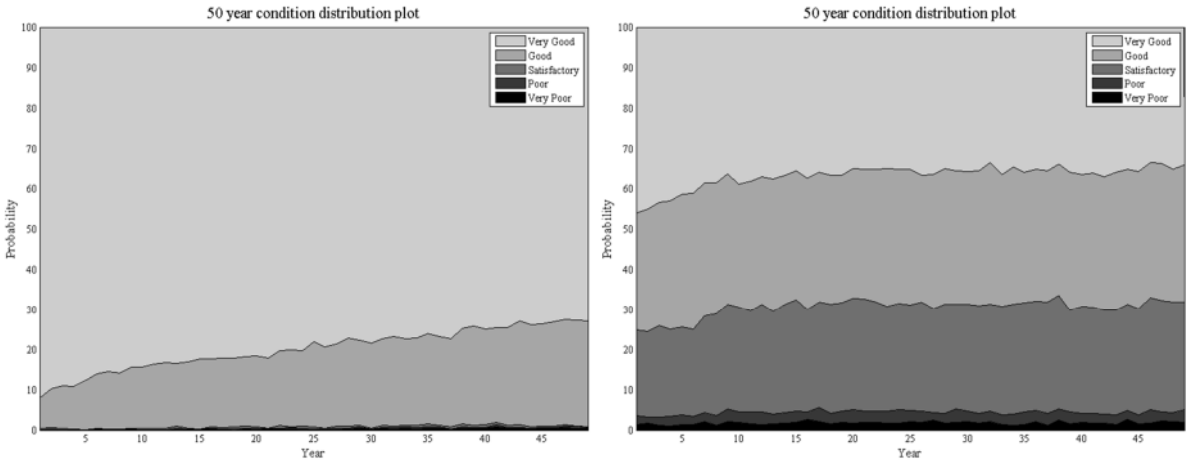


Figure 8. Simulation outputs



(a) (b)

Figure 9. 50 Year Condition Distribution Plots

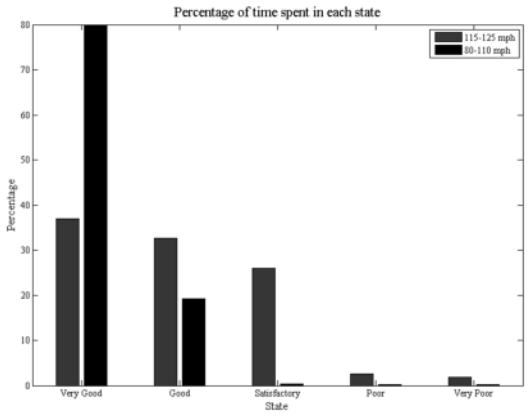


Figure 10. Percentage of Time Spent In Each State

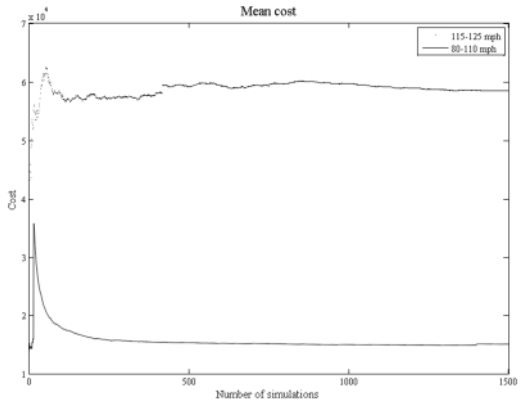


Figure 11. Mean cost

11. Conclusions

The genetic algorithm revealed a different asset management policy for each speed band. This adds weight to the argument that a one strategy for all is not the most cost effective solution. Furthermore, maintaining higher line speeds increases costs and increases downtime due to increased interim maintenance and renewals.

Acknowledgement

John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is also Director of The Lloyd's Register Foundation (LRF)[1] Centre for Risk and Reliability Engineering at the University of Nottingham. Matthew Audley is conducting a research project funded by Network Rail. They gratefully acknowledge the support of these organisations.

References

1. Andrews JD. Railway Track Ballast Condition Monitoring, Proc. of ESReDA Seminar: Advances in Reliability-Based Maintenance Policies, La Rochelle, France, 5-6 Oct 2011.
2. Prescott DR and Andrews JD. A Railway Track Ballast Maintenance and Inspection Model for Multiple Track Sections, In Proceedings of PSAM 11 (Probabilistic Safety Assessment and Management) / ESREL 2012 (European Safety and Reliability Conference), Helsinki, Finland, 25-29 June 2012.
3. Andrews JD. A Modelling Approach to Railway Track Ballast Asset Management, Proc. of the IMechE, Part F: J. Rail and Rapid Transit, first published on 13 July 2012. DOI: 10.1177/0954409712452235(2012).
4. Prescott DR and Andrews JD. A Track Ballast Maintenance and Inspection Model for a Rail Network, Proc. of the IMechE, Part F: J. Rail and Rapid Transit. (To be published in 2013).
5. Audley M and Andrews JD. The Effects of Tamping on Railway Track Geometry Degradation, Proc. of the IMechE, Part F: J. Rail and Rapid Transit. (To be published in 2013).

[1] Lloyd's Register Foundation supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

Asset Management of a Railway Signalling System

Raphaelle Barbier Saint Hilaire, Darren Prescott and John Andrews
Nottingham Transportation Engineering Centre,
University of Nottingham, Nottingham, UK

Abstract

Signalling systems are crucial to the safety, capacity and efficiency of the railway. Their development has been integral to the evolution of the railway, to the point that nowadays railway operations rely heavily on signalling systems. The large number of signalling-related, fatal accidents shows the importance of signalling to safety and the daily occurrence of delays in rail networks demonstrates the effect that signalling systems can have on rail services. Therefore, it is important that asset management programmes consider the reliability and safety of the signalling systems.

This work considers the use of coloured Petri nets to model the railway and study the impact of incidents on safety and delays. Models of each asset of the signalling system of a line section have been developed, as well as models of incidents that affect those assets. These sub-models are then connected in order to create a line model. The application of this model to an example line section is then used to demonstrate the results that can be obtained from such a model.

1. Introduction

Asset management is defined in PAS 55 [1] as the “systematic and coordinated activities and practices through which an organization optimally and sustainably manages its assets and asset systems, their associated performance, risks and expenditures over their life cycles for the purpose of achieving its organisational strategic plan”. There is particular interest in the application of asset management principles to the management of railways, where cost, performance and safety are of national significance. The asset management strategy can be varied by changing policies such as those relating to corrective or preventive maintenance, system or component renewal or the number of maintenance personnel. Asset management for the railway infrastructure places an emphasis on evidence-based decision making, using knowledge of how assets degrade and subsequently fail to attempt to optimise the timing of maintenance and renewal interventions. Its main objectives are to minimise whole life costs and ensure that the railway functions safely.

Over the last few decades, the railway industry has developed a holistic approach to analysing ‘whole life’ costs using the now commonly-used techniques of ‘Life-Cycle Costing’ (LCC) and ‘Whole Life-Cycle Costing’ (WLCC). WLCC is an economic assessment that considers all relevant cost flows over the life of an asset. It is used in option evaluation when procuring new assets and in decision-making to minimise whole-life costs throughout the life of an asset through maintenance and renewal operations [2-5]. WLCC

model is based on four principal asset relationships:

- asset age and utilisation to asset condition;
- asset condition to probability of failure;
- probability of failure to train performance and safety performance;
- impact of interventions on asset age and condition.

In general, the condition of signalling assets is assessed visually. Using the condition of every asset, their initial predicted nominal life and their usage models predict the actual nominal life and the risk of serious failure. There is little information in the literature about the relation between the probability of failure and the service output impacts and it appears that some WLCC models tend to be partly based on engineers' experience.

In this work, we present a method to investigate the impacts of incidents on delays and safety. Section 2 discusses the signalling system. Section 3 presents the Petri Net and Coloured Petri Net modelling methods. Section 4 presents models of the different assets of the signalling system as well as a method to model incidents. The connection of the sub-models to form a line model is presented in section 5. In section 6 the model is applied to a line in order to investigate the delays and safety associated with signalling system incidents. Finally, concluding remarks are given in Section 7.

2. Signalling system

A signal block is a section of line between two successive signals. Modern signalling systems aim to protect trains and enable them to operate on schedule by regulating their advance along the line, signal block by signal block. The assets making up the signalling system, train detectors, signals and protection devices, are presented in this section along with the related failures or incidents that are considered in this work. These failures and incidents have been chosen for their importance in causing service disruption or safety issues.

2.1 Signal

Signals were originally mechanical systems, and although these mechanical systems are still used in some places, they have for the most part been replaced by electrical coloured-light signals. These can have bulbs or LED matrices with up to four different aspects. In this paper, three-aspect signals are considered. Their aspect sequence is red (danger – stop) → yellow (caution – slow down) → green (clear – proceed at normal speed). The red aspect is displayed when there is a train in the signal block, the yellow aspect when this signal block is clear but the following one is occupied and the green aspect when this signal block and the following one are both clear.

2.2 Train detector

Train detectors are used to detect whether track sections are occupied and this information is then used in the control of the signals. The section of line monitored by one detector is referred to as a block. Different types of detector exist, the main ones being spot wheel detectors such as axle counters and

linear detectors such as track circuits. This paper focusses on axle counters. As a train enters a block, its axles pass a detection point and the counter increments, recording the number of axles. As the train exits the block, its axles pass another counting head and the counter decrements, resetting itself to zero.

Incident	Explanation	Repair actions
SPAD category A	A Signal Passed At Danger category A is a red signal correctly displayed is not respected by the driver.	No actions needed
Blown signal bulb	A blown yellow or green bulb leads to a fail-safe situation with a respectively red or yellow signal automatically displayed. A red blown bulb leads to an unsafe situation with a yellow signal displayed.	The bulb is changed and the signal is set to display a red signal. Then the normal sequence starts again and the adequate aspect can be displayed.
Electricity failure	No electricity supply due for example to damage to electric cables cables	Repair/replacement of the cables and reset of the signalling signal sequence.
Red signal not displayed	A red signal is not displayed when it should. It can be due to a an error from the signaller or an electronic error.	If due to electronic error, repair the system. Display a red signal and let the sequence start again.
Axle counter miscount	The number of axles is miscounted. Here the case of one too many axle detected will be considered.	The presence of a train in the block and its number of axles is studied and the counter is reset to its correct value.
Axle counter deficiency	A failure such as electrical failure resets the counter to zero.	Repairs of cause of the deficiency, check of number of axles in the block and reset of the counter to its correct value.
AWS activation failure	The AWS fails to get activated when a signal is red and the alarm is not set off when a train passes the AWS.	Checking of the functioning of the magnet and the magnet detection in the train. The failing parts are repaired and the system reset.
No reaction to AWS alarm	The AWS alarm is ignored by the driver.	Check on the driver. Take appropriate action such as changing driver.
Emergency stop failure	The AWS fails to stop the train when an emergency stop is performed.	Repair the automatic emergency brake.
Unexpected emergency stop	An error leading to an unnecessary activation of the emergency brakes caused by a failure in the automatic deactivation of AWS or in the deactivation by the driver.	Check on the cause and conduct appropriate maintenance actions. The system is then reset.

Table 1. Table of the considered signalling incidents and repair actions

2.3 Protection device

The train protection device can be a speed controller operating in a speed limit zone, or train stopper preventing a train from passing a signal at danger. The device considered in this paper is the British Automatic Warning System (AWS), used as a red signal protector and composed of two magnets. When a train passes an AWS, a permanent magnet initiates a warning tone. The second magnet is an electromagnet which will automatically reset the system if the signal is green or yellow. If the signal is red, the system has to be reset by the driver in less than 15 seconds, otherwise the train will be emergency braked.

2.4 Incidents

The incidents (and failures) affecting the operation of the signalling system considered in this work are caused by poor asset condition or human error. Another common cause is vandalism. The incidents, their consequences and the repair action undertaken are presented in Table 1.

3. Petri net method

3.1 Petri net

Petri nets were first defined by Carl Adam Petri in his PhD thesis “Kommunikation mit Automaten” [6] (“Communication with Automata”). Petri nets [7] are a graphical and mathematical modelling tool for the description and analysis of synchronisation, communication and resource sharing between concurrent processes.

A simple example of a Petri net is presented in Figure 1. Petri nets are based on four basic elements illustrated in the simple example in Figure 1. The places (circular nodes) describe the possible system states. Transitions (rectangular nodes) represent the actions leading to a change of state. Arcs (arrows) are orientated connections between a transition and its input places (representing the necessary state for an action to happen) and its output places (representing the state once the action has been executed). Finally, places can contain tokens (black dots) which describe the current state of the system.

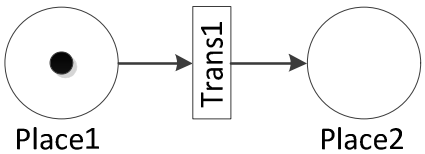


Figure 1. Simple example of a Petri net

A transition is enabled if each of its input places contains at least one token. In Figure 1, Place1 has a token so Trans1 is enabled. When a transition fires it consumes a token from each input place and produces a token for each

output place, creating a new state of the system. On Figure 1, the token in Place1 would disappear and a token would appear in Place2 as Trans1 fires. Trans1 would not be enabled anymore. Stochastic Petri nets include a notion of time with transitions being able to carry time delays.

3.2 Coloured Petri net

Coloured Petri nets (CPN) were first defined by Kurt Jensen [8]. They are a high-level Petri net language developed to simplify large models. They incorporate both data structuring and hierarchical decomposition and therefore are well-suited for systems consisting of a number of assets which communicate and synchronise.

The innovation in CPN is the implementation of “colours” which are data carried by tokens. Each place has a colourset assigned to it and can only receive tokens of that colourset. In Figure 2, Place1 (of colourset INTEGER) contains two tokens of respective colours ‘2’ and ‘5’ and timestamp 3 and 0. Input arcs either carry global variables (in Figure 2, integer variable ‘n’) or specific data. Transitions can carry guard conditions on the values of input tokens (in Figure 2, [n<4]) and a timestamp for the output tokens (in Figure 2, time of firing + 7). Output arcs can carry either specified values (in Figure 2, the boolean value ‘true’) or expressions computed using the input variables. In Figure 2, Trans1 will only be enabled at time 3 as only the token of colour ‘2’ respects the guard. At time 3, this token will be consumed and a token of colour ‘true’ will be sent to Place2 (of colourset BOOLEAN) with a timestamp of 3+7=10.

Kurt Jensen is also the founder of the CPN Group at Aarhus University, Denmark, who developed CPN Tools, a tool for editing, simulating, and analysing coloured Petri nets which was used for the modelling, simulation and data recording in this work.

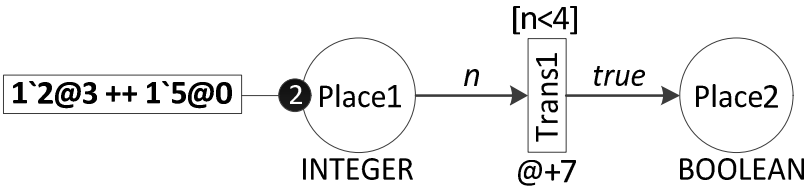


Figure 2. Simple example of a Coloured Petri net

4. Signalling asset models

A coloured PN model of a line section has been created using CPNTools. It contains modules that account for: the behaviour of the line and trains that use it, the axle counters that detect the presence of trains along the line, the signals and the automatic warning system. These modules are described in the following subsections. For clarity, the structure of the PN is presented for each module and the features are explained in the accompanying text.

4.1 Train generator and line section

A section of line containing a number of consecutive signal blocks has been modelled. Each block is represented by one transition ‘i’ and two places ‘Block i’ (which can contain tokens with a train label), and ‘List i’ (which contains a unique token with a list label ordering the trains as they entered the block). When transition ‘i’ fires, it consumes tokens from the previous line section and sends a token to both places. The information attached to the train token when it was first generated are conserved and a delay is added to its timestamp to represent the expected time spent by the train in the block. At the end of the line, the ‘Line Exit’ place records all the trains which passed along that section of line.

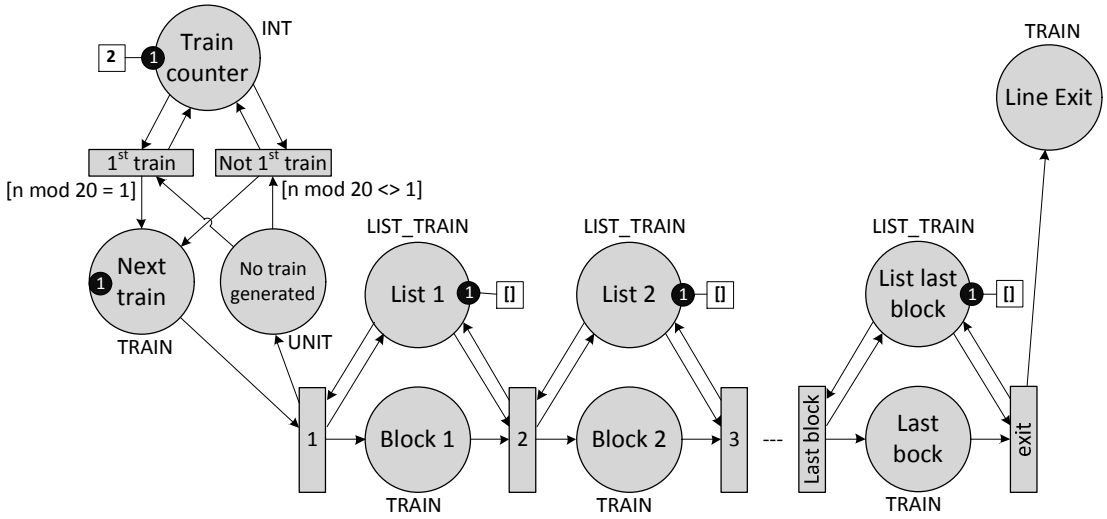


Figure 3. Model of the train generator and the line

The train generator is composed of three places. ‘Next train’ contains the token whose colour details the information on the next train to enter the line (train name, train type, number of axles, time at which the train will enter the line and planned time of arrival at the stations). The firing of transition 1 represents the entrance of the train in the first section of line ‘Block 1’. It moves the train token from ‘Next train’ to Block 1 and adds the train to the list carried by the token in List 1. It also sends a token to place ‘Need train’, which means a new train needs to be generated. However the delay between the time the last train entered the line and the time the next train will be allowed to enter the line varies depending on whether this new train is planned on the same day as the last train or will be the first train of the following day. This is why two transitions with different delay guard are needed (n is the colour of the token in ‘Train counter’, “mod” is the modulus function):

- Transition ‘1st train’ has a guard ‘[n mod 20 = 1]’, and requires the colour of the token in every ‘List’ place to be empty lists to be enabled.
- Transition ‘Not 1st train’ has a guard ‘[n mod 20 <> 1]’, and a constant delay of 30 minutes.

The firing of one of the two transition consumes the token in ‘Need train’, increment the token in ‘Train counter’ and generates a new train in ‘Next train’.

4.2 Axle counters

The model for axle counters is composed of a place ‘AC’ per axle counter. It contains a single integer token initially labelled 0. When a train is present in a block, such as block i-1 in Figure 4, the value of the counter is equal to the number of axles on that train, such as 10 in ‘AC i-1’. When transition i fires (as described in Section 4.1), it reads the number of axles of the train in the label of the train token, subtracts that number from the value in ‘AC i-1’ and adds it to the one in ‘AC i’.

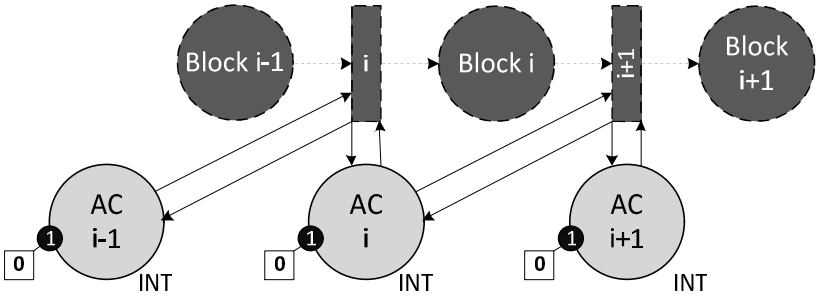


Figure 4. Axle counter model

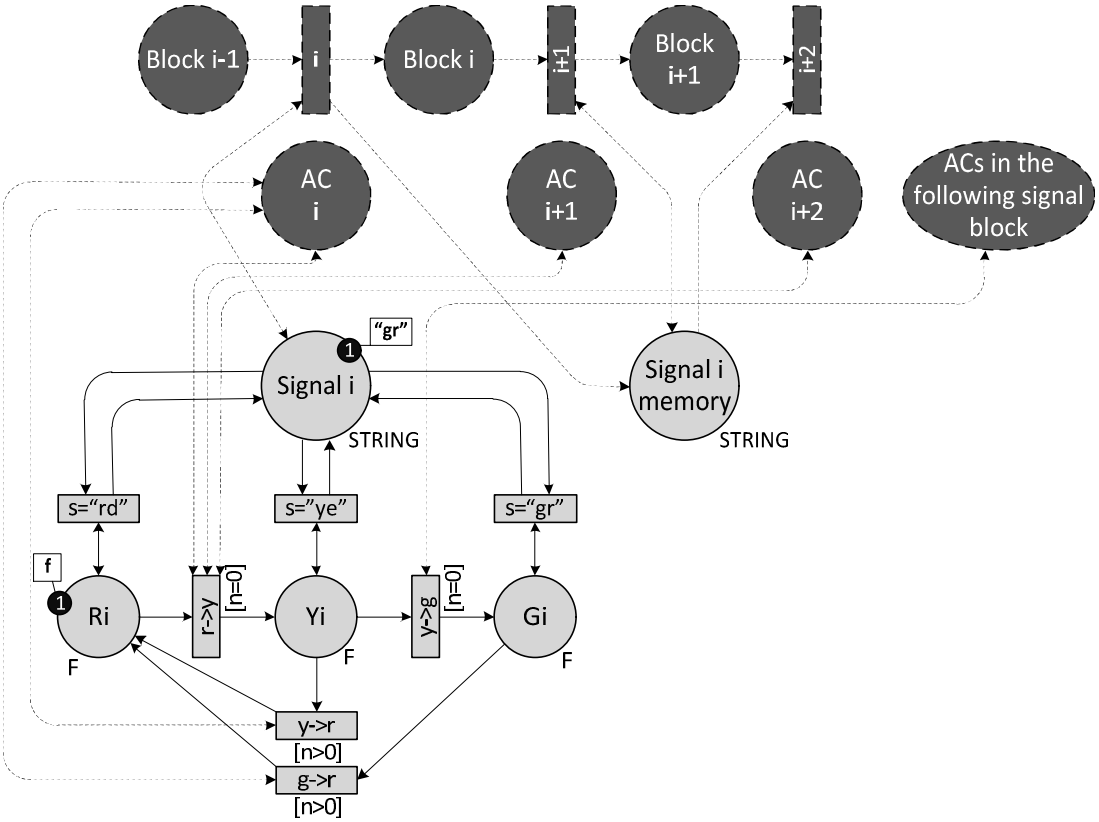


Figure 5. Three-aspect signal model

4.3 Three-aspect signal

The model of a three-aspect electrical signal presented in Figure 5 contains five places. Places ‘Gi’, ‘Yi’, ‘Ri’ represent the three aspects green, yellow and

red. The one with a token is the displayed aspect. Four transitions model the change of aspect. As soon as a train enters block i, either ‘g->r’ or ‘y->r’ fires, turning the signal to red. For ‘r->y’ to fire, the values of the tokens in all ‘AC’ places in the signal block have to be 0. For ‘y->g’ to fire, the values of the tokens in all ‘AC’ places in the following signal block also have to be 0. Place ‘Signal’ contains a string token with the displayed aspect as label. Finally, ‘Signal i memory’ contains – when there is a train in the signal block – a token with the aspect of the signal when the train passed it. Those places are used for the connection between the signal sub-model and the sub-models representing the line sections and the axle counters.

4.4 Automatic Warning Signal

The model of an automatic warning system (AWS) presented in Figure 6 contains four main places and two timer places. When the signal turns red, the AWS is activated and t1 will fire when a train approaches the signal. The value of the token in ‘Alarm rings’ is changed to the Boolean value ‘yes’ and the token in ‘timer 1’ is given a timestamp of time+15. If in the next 15 seconds the value of the token in ‘Driver reacts’ is ‘yes’, t2 fires, switching off the alarm. Otherwise t3 fires setting off the emergency brakes. t4, ‘timer 2’ and t5 allow the reset of the value of the token in ‘Emerg. stop’ to ‘no’, stopping the emergency brake procedure and releasing the train.

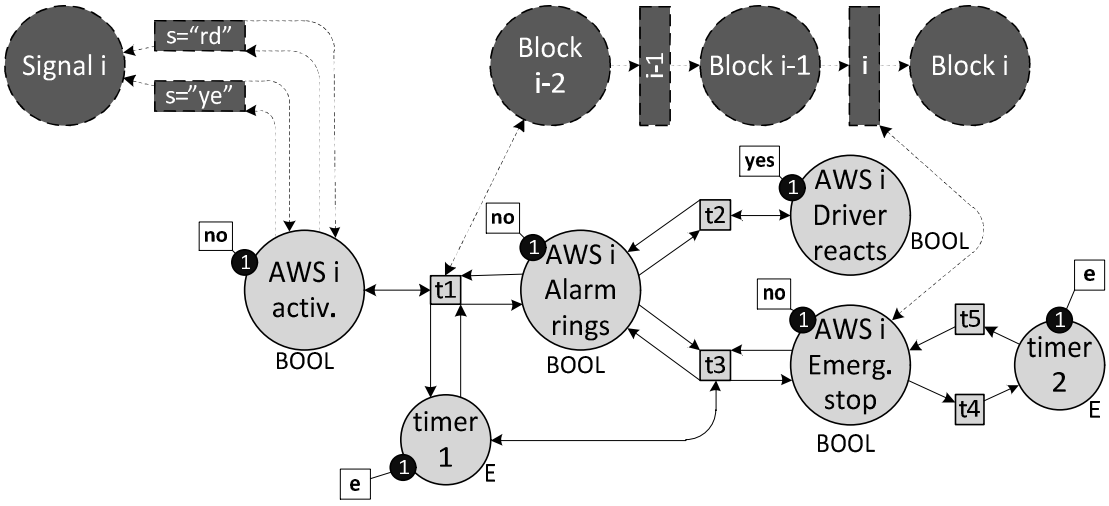


Figure 6. Automatic Warning System model

4.5 Incidents

The failure and incident models follow the same pattern. The model of axle counter deficiency is presented in Figure 7 as an example.

One place with the name of the incident (‘AC i deficiency’) contains a timed token defining when this incident will happen next. The transition with the incident name (‘AC_Def’) sets off the effects of the incidents, sending a token to the ‘repair’ place and one to the ‘Distrib.’ place.

The token in place ‘repair’ is given a time stamp corresponding to the delay between the incident and the repairs/maintenance actions. When this token becomes available, the repair actions (transition ‘Repair’) are conducted. For some incidents, the ‘Repair’ transition also requires inputs from the rest of the model (here from place ‘List i’ to obtain the number of axles in the block).

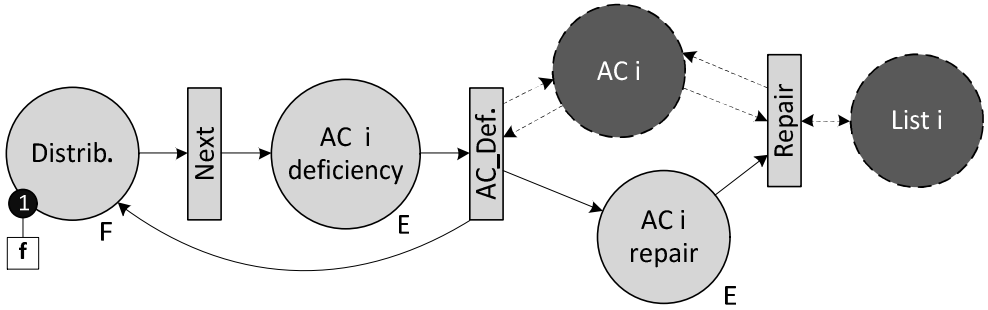


Figure 7. Models of miscount and deficiency of axle counters

The token in the ‘Distrib.’ place triggers transition ‘Next’, firing tokens into an accident place and calculating its time stamp using an exponential distribution. The models of the incidents presented in Table 1 have the characteristics presented in Table 2.

Incident	Incident output	Repair input
SPAD category A	block-to-block transition: guard signal = red	no repair
Blown signal bulb	token moved from blown aspect place to replacement aspect place	no input,aspect sequence reset
Electricity failure	token in aspect places consumed	no input,aspect sequence reset
Red signal not displayed	token moved from red to green aspect place	no input,aspect sequence reset
Axle counter miscount	place 'AC': value = value +1	place 'List i' for the number of axles in the block
Axle counter deficiency	place 'AC': value = 0	place 'List i' for the number of axles in the block
AWS activation failure	place 'AWS Activ': boolean 'no'	no input, aws boolean values reset
No reaction to AWS alarm	place 'AWS Driver reacts': boolean 'no'	no input, aws boolean values reset
Emergency stop failure	place 'AWS Emerg. stop': boolean 'no'	no input, aws boolean values reset
Unexpected emergency stop	place 'AWS Emerg. stop': boolean 'yes'	no input, aws boolean values reset

Table 2. Table of the characteristics of the incident models

5 WHOLE SYSTEM MODEL

To form a model of a section of line, the sub-models described in Section 4 have to be connected. Several arcs are needed to link the different sub-models and represent the interaction existing between the different assets. The arcs connecting the sub-models are represented by dotted arrows in the figures in Section 4.

5.1 Line Section and axle counters

Line sections are connected as shown in Figure 3: each transition modelling the entrance of a train into a block becomes the transition modelling the exit of a train from the previous block. As they are connected, the axle counters can be added to the model. For each block i, the axle counter ACi is connected to transition i to increment the number of axles entering block i and to transition i+1 to decrement the number of axles leaving block i.

5.2 Signal

The signal sub-models are connected to the line section sub-models. A signal i regulating the entrance of a train into block i is connected to transition ‘i’. A guard prevents the firing of transition ‘i’ if the value of the token in place ‘Signal i’ is ‘rd’ (red). The delay added to the timestamp of the train token is calculated on the length of the line section but also on the aspect displayed: a ‘ye’ token in ‘Signal i’ means the train will slow down so the delay is doubled. For each section of line not starting with a signal, the delay is calculated on the colour of the signal when the train passed it, that is to say the colour of the token in place ‘Signal i memory’.

The signal sub-models are also connected to the axle counter models: the colours of the tokens in places ‘AC' have to satisfy the guards regulating the firing of transitions that change the displayed aspect.

5.3 AWS

The AWS sub-models are connected to the signal sub-models for the activation of the AWS (Boolean ‘yes’ in place ‘AWS Activ.’ when the signal turns red and ‘no’ when it turns yellow). It is also connected to the line section sub-models for the triggering of the alarm when a train approaches the AWS (‘yes’ in place ‘AWS Alarm rings’ when transition ‘i-2’ fires). Finally transition ‘i’ can’t fire if an emergency stop was done (‘yes’ in ‘AWS Emerg. stop’).

5.4 Inputs

Three main inputs can be varied in this model:

- the traffic density, by varying the number of trains per day and their separations in minutes.
- the incident distribution of each asset, by varying the type of distribution or the parameters of the chosen distribution.
- the repair distribution of each asset by varying the type of distribution or the average delay of the chosen distribution. This allows different corrective maintenance strategies to be modelled.

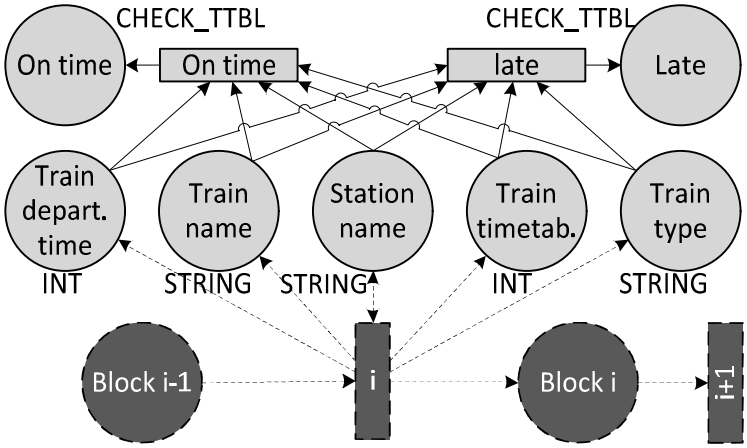


Figure 8. Model of the delay recorder

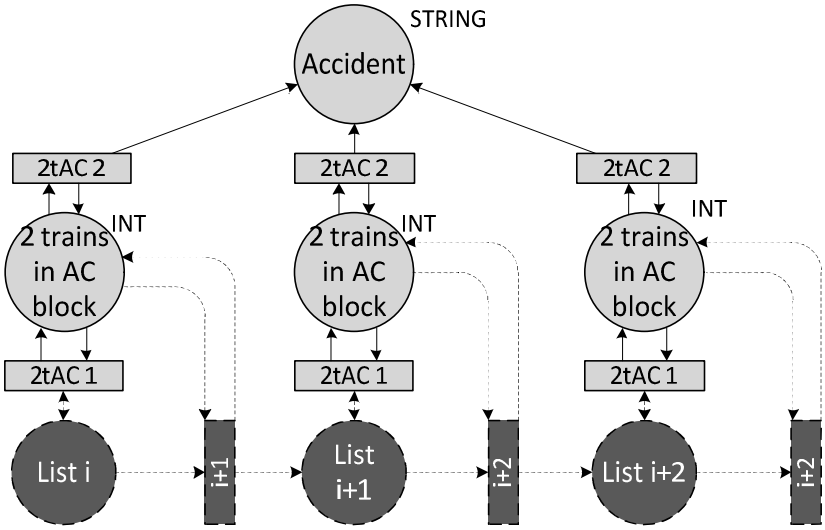


Figure 9. Model of the accident recorder

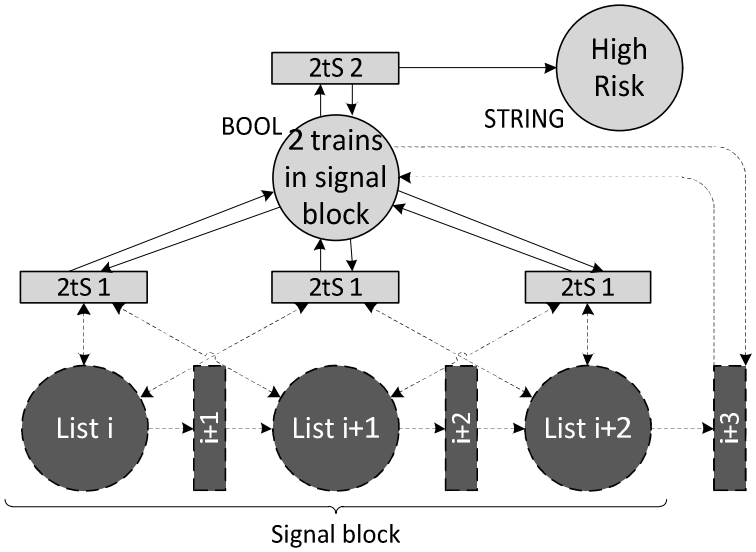


Figure 10. Model of the high risk recorder

5.5 Outputs

Two categories of outputs have been implemented: train delays and safety risk. Train delays are measured every time a train enters a station (firing of transition ‘i’ when there is station in block i). On the model in Figure 8, the firing of transition ‘i’ extracts information from the train token. The difference between the simulation time and the train timetable is submitted to guards on transitions ‘On time’ and ‘Late’. One of the two transitions will fire, sending a token to place ‘On time’ or ‘Late’ with a record of the information on the train, the station and the delay as colour.

The second output is related to safety. Different levels of risk can be studied but two will be presented in this section:

- Risk of accident occurrence (two trains in the same axle counter block) a model of which is presented in Figure 9. When the length of the list in place ‘List i’ is bigger than 1, transition ‘2tAC 1’ can fire, changing the value of the token in ‘2 trains in AC block’ from ‘0’ to ‘1’. ‘2tAC 2’ can fire, changing the ‘1’ to a ‘2’ and sending a token to the ‘Accident’ place with a record of the trains and the simulation time as colour. The ‘2’ is changed back to ‘0’ when transition ‘i’ fires (i.e. when one of the two trains leaves the block).
- High risk of accident occurrence (two trains in the same signal block), a model of which is presented on Figure 10. When the length of the lists in the two places ‘List i ’ and ‘List i+1’ are bigger than 1, transition ‘2tS 1’ fires, changing the value of the token in ‘2 trains in signal block’ from ‘0’ to ‘1’. ‘2tS 2’ can fire, changing the ‘1’ to a ‘2’ and sending a token to the ‘High Risk’ place with a record of the train and the simulation time as colour. The ‘2’ is only changed back to ‘0’ when the transition symbolising the train exiting the signal block fires.

6 APPLICATION

In this section the model is applied to an example line section and the results that can be obtained from the model are demonstrated.

6.1 Description of the model

The model has been applied to a line ten miles long and passing through 6 stations. Its signalling system is composed of 26 blocks delimited by 27 axle counters, 9 three-aspect signals associated with 9 Automatic Warning Systems. The model of the line is composed of the sub-models presented in Section 4 and connected using the technique discussed in Section 5.

6.2 Inputs

A set of 220 simulations was run for 20 trains a day, 5 days a week, for a year, with a train every 30 minutes and a delay of one hour for the repair actions. The incident distributions modelled were exponential distributions with mean time between failures (MTBF) presented in Table 3. These values are

representative of MTBF that might be seen in reality. Lack of data on corrective maintenance distributions meant that constant delays were chosen for repairs; these are presented in the same table.

Asset	Incident	MTBF (yrs)	Repair (min)
Signal	SPAD category A	10	-
Signal	blown bulb	8	60
Signal	electrical failure	5	60
Signal	red signal not displayed	10	10
Axle counter	Miscount	2	60
Axle counter	Deficiency	2	60
AWS	activation failure	3	60
AWS	no driver reaction	5	20
AWS	emergency stop failure	3	60
AWS	unexpected emergency stop	8	60

Table 3. Table of the values used for incident and repair distributions

6.3 Outputs of the model

This study aims to increase the understanding of the impact of incidents on delays and safety. Therefore the considered outputs were the number of trains arriving late at stations, the number of times when two trains were in the same axle counter block (indicating a risk of accident occurrence), and finally the number of high risk situations when two trains were in the same signal block indicating a lower risk of accident occurrence.

6.4 Results

Figure 11 proves that a sufficient number of simulations were run to obtain representative values. This figure also provides the average of the outputs: the average of number of accidents is 0.7 accidents per simulation, and the average number of delayed trains is 100 trains per simulation. The high average number of delayed trains is representative of the well-known issue of delays on the railways; it shows that for the modelled line section the modelled incidents play an important role in the delays. Figure 12 presents histograms of the three outputs. A difference of shape between the safety output histogram and the delay output can be noticed. The number of delayed trains per simulation varies from 30 to 200 with a maximum occurrence of 110 delayed trains. The high variations in the number of delayed trains indicate the dependence of this value on the incidents that occur during the simulated period. Implementing changes in preventive maintenance in order to reduce the mean time between failures or in corrective maintenance to be able to restore traffic on a line quicker could be an efficient way to reduce the impact of those incidents and hence the delays.

As seen in Figure 12, there are no situations relating to low or high risk of accident in around half of the simulations performed. The remaining simulations saw between 1 and 4 low risk or high risk situations.

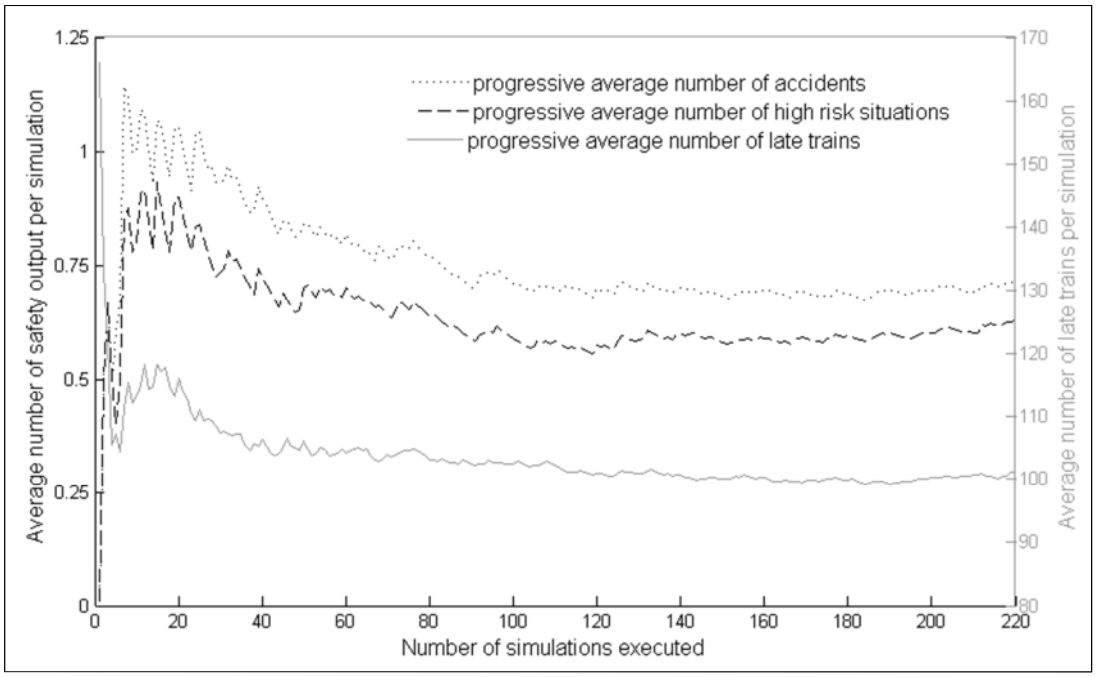


Figure 11. Cumulative average

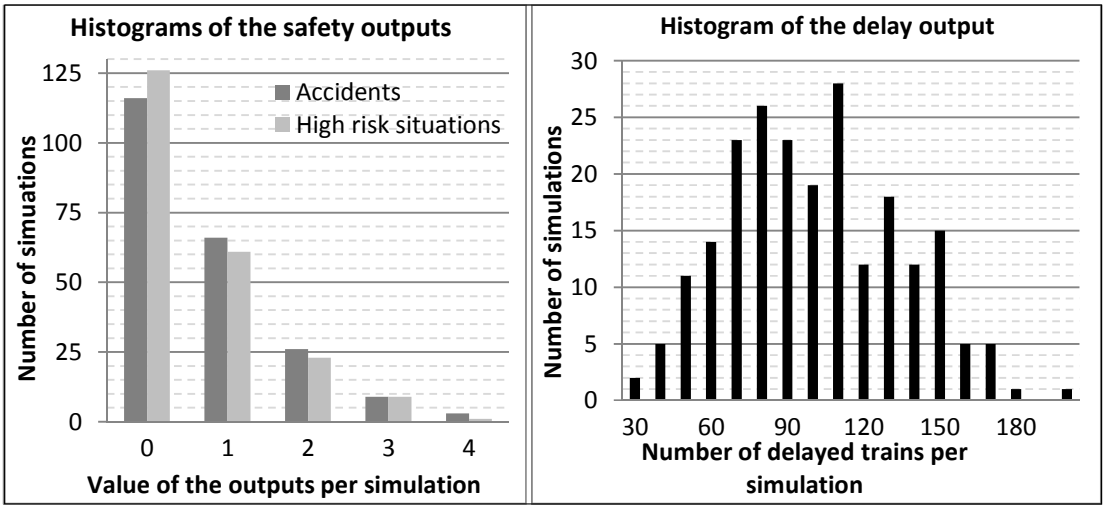


Figure 12. Results of the study

7 CONCLUSION

This paper presented a coloured Petri net model that can be used to investigate the asset management of railway signalling systems on the railway network. The model consists of modules representing the different assets in the signalling system, which are connected to form a mode of an entire line. The model offers the possibility to vary the traffic density and incident and repair distributions. The model produces outputs that allow the analysis of situations affecting safety and delays, allowing these factors to be considered when making decisions during the asset management process. The model was applied to an example line section and used to demonstrate the results that can be obtained: the number of situations where two trains are in a signal block (low risk of accident), the number of situations where two trains are in

an axle counter block (high risk of accident) and the number of delayed trains on a line. The model could also be used to analyse the distribution of trains per delays. The model was analysed using Monte Carlo simulation and could be used to investigate the effect of changes to either the asset management strategy or the assets on the line. Results can be compared to decide if the benefits in term of increased safety and reduced delays are worth the expense of the alternative under investigation.

Acknowledgement

John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is also Director of The Lloyd's Register Foundation (LRF)[1] Centre for Risk and Reliability Engineering at the University of Nottingham. Darren Prescott is the LRF Lecturer in Risk and Reliability and is based in the Centre for Risk and Reliability Engineering at the University of Nottingham. Raphaelle Barbier Saint Hilaire is conducting a research project funded by Network Rail. They gratefully acknowledge the support of these organisations.

References

1. Institute of Asset Management, British Standard Institution, *PAS55-1:2008 Asset Management Part : Specification for the optimized management of physical asset,* 1st ed., BSI (September 2008)
2. Skinner, M., Kirwan, A., and Williams, J. *Challenges of Developing Whole Life Cycle Cost Models for Network Rail's Top 30 Assets*, Asset Management Conference 2011, IET and IAM , vol., no., pp.1,6, (30th Nov. 2011- 1st Dec. 2011)
3. Lackhove, C., *et al*. *Advancing life-cycle-management for railway signalling and control systems*. *In:* FINKBEINER, Matthias, ed. *Towards Life Cycle Sustainability Management - LCM 2011*, Berlin, 28 August – 01 September, 2011.
4. Brownless, G., Turner, S. and HSE. *An Asset Management Model for UK Railway Safety – Literature Review and Discussion Document*. Buxton: Health and Safety Laboratory (2005)
5. Boussabaine, H. A., Kirkham, R. J. *Whole Life-cycle Costing Risk and Risk Responses*. Oxford: Blackwell Publishing. (2008)
6. Petri, C. A. *Kommunikation mit Automaten*. Ph. D. Thesis. University of Bonn (1962)
7. Murata, Tadao. Fellow, IEEE, *Petri Nets; Properties, Analysis and Applications*, Proceeding of the IEEE, Vol. 77, No. 4, (April 1989)
8. Jensen, K. *Coloured Petri nets : basic concepts, analysis methods, and practical use,* 2nd ed., Berlin : Springer (c1996-1997)

[1] Lloyd's Register Foundation supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

Modelling Railway Service Reliability

Claudia Fecarotti, John Andrews and Rasa Remenyte-Prescott
Nottingham Transportation Engineering Centre, University of Nottingham, Nottingham, UK

Key words: service reliability, flexibility, redundancy, discrete-event simulation.

Abstract

Current forecasts suggest that by the 2020s, the level of traffic on the main UK railway lines will challenge the capacity of the system to a degree which requires a more radical solution than small incremental changes to the railway system. In order to avoid massive congestion of the railway network, the use of the current infrastructure need to be optimised. In this context the availability of the railway network 24 hours a day, 7 days a week has been stated as a potential solution to meet capacity requirements. However a greater utilisation of the existing infrastructure will accelerate degradation processes and reduce the time slots available for maintenance. Therefore the system needs to be designed, operated and maintained in a significantly different way if long disruptions to the service are to be avoided resulting from system failures.
The fundamental idea of this research is that network flexibility is a key feature to improve the ability of the system to cope with disturbances during operation, thus providing a more reliable service. In this paper a methodology to explore different possibilities to increase network flexibility in order to use the infrastructure more efficiently and effectively is presented. A discrete-event simulation model has been developed to simulate railway operation and assess service reliability. The model will provide an assessment of network stability for different system designs and perturbed scenarios. In the paper, rail operation in a section of the UK network is simulated for a number of infrastructure and failure scenarios and results are discussed.

1. Introduction

The increasing demand of railway transport in terms of both passengers and freight is a challenge for the network operators. Planning significant changes of the railway network, such as additional lines or platforms at stations, imply excessive costs and are often subjected to physical constraints imposed by the current infrastructure layout, such as land availability and restrictions of the civil structures including bridges and tunnels. Therefore the availability of the railway network 24 hours a day, 7 days a week has already been stated as a possible solution to meet the capacity requirements. Running a 24/7 railway service would imply a greater utilisation of the infrastructure and at the same time reduce the time slots for maintenance. This would determine higher deteriorations rate of the railway asset, and therefore higher probability of failure, thus making the railway system more susceptible to disruptions. Furthermore, increasing capacity by maximising the use of the available infrastructure requires fitting more services and operating the network at its

capacity margins, thus affecting system performances. Common approaches to improve service reliability, such as timetable robustness, are not enough to handle with disruptions especially when large disturbances occur. Increasing capacity whilst providing high level of service reliability is a challenge for the rail network operators.

The basic idea of this paper is that flexibility is a fundamental property of the network to increase capacity without threatening the service reliability. Flexibility is essential to avoid long disruptions to the service resulting from technical failures or any other external cause. Switches are the main elements to provide flexibility since they enable trains to be guided from one track to another thus allowing crossing and overtaking operations. The number and location of switches in the railroad network may affect significantly the ability of the system to cope with disturbances. In this paper a methodology for the systematic evaluation of the effects of additional switches on service reliability is presented. Railway service reliability is evaluated not only with respect to small disturbances but also major disruptions which cannot be handled by simply improving timetable robustness. A discrete-event simulation model has been developed as a tool to simulate railway operation and assess system performance for different potential infrastructure designs and disrupted scenarios. Potential system failures have been preliminarily analysed according to the FMEA approach. The analysis results in a database where system components are detailed and qualified in terms of their functions, their failure modes and corresponding effects on system performance, their severity and failure rates as well as the compensating actions to maintain the system safe. This list of failures becomes the source from which failures are randomly generated according to their probability of occurrence and introduced into the system during the simulation.
This paper is organised as follow. In section 1 motivations for this work are provided, followed by a literature overview on railway simulation in section 2. In section 3 system performance and measures for service reliability are discussed. The failure modelling approach is introduced in section 4, followed by a description of the proposed simulation model in section 5. Numerical simulations and conclusions are provided in section 6 and 7 respectively.

2. Literature overview

Railway service reliability is a complex matter since many factors affect the railway operations, some of them are related to system design, system reliability and maintainability, other concerns with the operational strategies and the planned usage of the infrastructure (timetables). Understanding and increasing service reliability is the objective of many researchers. In the literature a common approach to study the reliability of railway services consists in investigating timetable robustness. A timetable is considered to be robust if it is able to handle small disturbances and readily absorb small delays. In this sense it is directly related to service reliability. Both analytical and simulation methods are applied for evaluating the effect of delays and timetable robustness. Simulation in particular is a valuable tool used to support the decisional process in design activities and traffic management. Several simulation models have been developed for both commercial and

research purposes. Among the simulation tools commercially available is worth mentioning NEMO and SIMONE, developed by ProRail in the Netherlands and used to analyse complex and large scale train networks. OpenTrack® and RailSys® are microscopic models developed respectively by the ETH Zurich and Leibniz Universitat Hannover. Other software are VISION and TRAIL. Railway simulation has been used for timetable construction and timetable robustness assessment (Carey and Carville, 2000, Salido et al., 2012). In Salido et al. (2012) analytical and simulation methods to measure timetable robustness in a single railway line are evaluated. The authors point out the trade-off between capacity and robustness, and therefore service reliability. As a matter of fact, the greater the used capacity, the higher is the risk of knock-on delays, due to small headways and buffer times. The price for robustness and service reliability is a loss of capacity and timetable optimality. The average speed is considered to be an important factor for timetable robustness, as well as traffic heterogeneity. Vromans et al. (2006) investigate the impact of timetabling principles and in particular timetable heterogeneity on service reliability. The attempt of the authors is to create more homogeneous timetable by reducing running time differences per track sections, thus decreasing interdependencies among trains and therefore delay propagation. Simulation is used in order to analyse and compare different timetables. Delay resistant periodic timetables and recoverable robust timetable are considered in Liebchen et al. (2010) and Di Stefano et al. (2011) respectively. This literature is mainly focused on the evaluation of the influence of timetable properties on service reliability, and it aims at improving timetable robustness. Railway simulation is also applied to assess network stability, as well as system capacity (Abril, et al., 2008). Middelkoop and Bouwman (Middelkoop and Bouwman, 2001) describe the architecture and potentials of SIMONE and apply the model to test the effects of infrastructure and operational changes in some critical sections of the Dutch railway network, on the timetable and network stability (Middelkoop and Bouwman, 2002). Dicembre and Ricci (Dicembre and Ricci, 2011) use OpenTrack® to investigate the correlation among capacity, block section length, timetable and operational program for high density lines, in particular urban corridors, under regular and perturbed scenarios. OpenTrack® was also used by Luethi et al. (Luethi et al., 2006) to test the potential benefits of a new real traffic management system and train control on the system stability and capacity. Application of RailSys® to a section of the German railway network in order to improve the timetable of the system is described by Demitz (Demitz et al., 2004). Simulation has been also applied to address more specific issues like optimisation of the signalling system layout to improve the capacity of the system (Gill and Goodman, 1992, Hill and Bond, 1995).

To the best of the authors' knowledge, research on the impact of network flexibility on railway service reliability is scarce. The influence of network topology on system availability and service reliability is usually qualitatively discussed, but rarely it is systematically investigated and assessed. Furthermore, the concept of fault tolerance of the railway network as a whole is barely considered. Only reliability analysis of sub-system is performed, such as signalling system (Panja and Ray, 2009) and power supply system (Ho and Mao, 2007). When simulation is applied for the evaluation of system

performances under disrupted scenarios, disruptions are simulated by introducing randomly delays to trains. Delays usually act on running times or dwell times at stations. Technical failures are not considered with their specific characteristics during simulation. This prevents the analysis from taking into account the connection between system performance and system failures. Failures affect the system differently depending on the equipment involved, on the failure rates and time to repair, and on the kind of compensating action to be taken to maintain the system safe. Therefore, in order to correctly evaluate the system performance, it is important to understand the system behaviour in terms of its failure modes and the corresponding effects. To this aim a FMECA approach is suggested as a useful tool to analyse the railway system in terms of its failure modes.

3. System Performance

The main function of the railway system is the safe transport of passengers and goods at the scheduled time. The system needs to be managed during its complete life cycle in order to meet the traffic demand and provide an available service in a cost-effective manner. A common indicator of the system performance level is the “system availability”, defined as the ability of the system to provide the safe transport of people and goods under given conditions and over a given time interval, assuming that the required source of help are provided (interlocking between points and signals, track clear detection, train separation)(EN 50126, CENELEC 1999). System availability mainly concerns with the probability that the service will operate according to timetables, hence it is very often referred to as “service reliability”. The latter needs to be distinguished from the “system reliability” that is related to components/system failures. System reliability is the ability of the system to operate without failure for a stated period of time under specified conditions. It concerns with the frequency of components/system failures which prevent the system from providing its required functions, and in turn, is related to the maintainability of the system. Therefore, system reliability and maintainability are fundamental factors for availability since any failure of technical component results in a disruption of the railway operation and a delay to the service. To achieve high level of reliability it is not enough to improve failure rates of single components or sub-systems, but it is necessary to introduce some form of fault tolerance into the system in order to assure the continuity of service even in case of failure. In particular the network layout affects operational and maintenance strategies to minimise the effects of technical failures on system availability.

High availability is an important requirement of the railway system. The increasingly and intensive usage of the railway infrastructure has made the system more susceptible to disruptions. As a consequence, delays affect the daily operation resulting in reduced service reliability. In order to improve system availability and provide a reliable service, it is important to reduce both primary and secondary delays. Primary delays are caused directly by malfunctioning infrastructure or external causes, but not by other trains. While secondary delays, also referred to as knock-on delays, are caused by earlier delays of other trains and occur because of the shared use of the

infrastructure and interdependencies among trains. One way to improve service reliability is to develop timetables robust enough to absorb primary delays and cause as small secondary delays as possible. Robust timetables are a good solution to cope with small initial delays, but in order to handle with major disruptions more drastic solutions need to be taken which involve on-line re-scheduling and re-routing trains. Many efforts have been made in the literature in order to develop sophisticated re-scheduling and re-routing techniques, but a fundamental condition for the railway system to cope with disturbances is its own flexibility.

3.1 *Service reliability measure*

Two aggregate measures are commonly used in order to evaluate railway service reliability, the total delay time and the observed punctuality.

In this paper, system performance are measured in terms of total delay time defined as the total amount of delay experienced by all the trains running during the simulation period and it is given by

$$D = \sum_{\substack{i=1 \\ s=1}}^{I,S} \left(t_{is} - t_{is}^{d} \right), \qquad (1)$$

where t_{is} is the scheduled arrival time of train i at station s, and t_{is}^{d} is the actual arrival time of the considered train at the same station. I and S are the number of trains and the number of stations respectively.

4. Failure modelling: FMECA technique

Failure Mode and Effects Criticality Analysis (FMECA) has been adopted to analyse potential failure modes of the railway system and determine their effects on railway operation. FMECA is a step-by-step procedure which consists of successively breaking the system into its sub-systems and components down to a level which depends on the information available and on the purpose of the analysis. To systematically analyse the system and its parts, a set of FMECA worksheets are usually compiled where for each item, failure modes and the corresponding effects at both local and system level are specified. In the railway system, failures affect network availability differently depending on their severity. Therefore failures may imply sections of track being out of service, or simply speed restrictions. The corresponding effects are disruptions to the railway operation such as delays or deleted journeys, which in turn affect the overall system performance.
When filling FMECA worksheets, failure modes are classified in terms of severity, giving information on their criticality on system performance. Detection methods and frequencies to detection can be considered as well as the compensating actions to assure safety conditions even in a degraded operation mode. Moreover failure rates and time to repair are detailed.
FMECA worksheets can become a proper database of potential failures to be used into the simulation model described in section 4. Each worksheet contains the information necessary to simulate a disruption. Failures are

generated according to their probability of occurrence, while the time to repair determines how long it will take to repair the failure. Furthermore the block section affected by the failure is specified as well as the corresponding compensating action to keep the system safe. Compensating actions detailed into the worksheet correspond to specific strategies implemented into the simulation model. Table 1 shows an example of FMECA worksheet for track circuit.

Identification	**Function**	**Failure Modes**
Track Circuit	Train detection	(1)Relay failure (2)Poor shunting (3)Foreign current

System effect	**Block section**	**Detection methods**	**Mean time to detect**
(1)Trains on track are not detected	15	(1)Condition monitoring	(1) d1
(2)Trains on track are not detected		(2)Condition monitoring	(2) d2
(3)Trains on track are not detected		(3)Condition monitoring	(3) d3

Severity	**Failure rate**	**Time to repair**	**Compensating action**
(1) high	(1) λ_1	(1) τ_1	Section closing
(2) high	(2) λ_2	(2) τ_2	
(3) high	(3) λ_3	(3) τ_3	

Table 1 Example of FMECA worksheet for track circuits.

Track circuits are the most used technology for direct detection of vehicles on tracks. For track circuits different failure modes can be identified which can be grouped into two main classes, “right-side” and “wrong-side” failures. In particular wrong-side failures lead to an unsafe state so they are usually classified with a high severity range. To this group belong all failures which prevent the circuit from detecting a train on the track, thus involving a risk of collision. If such a failure occurs the traffic on the affected section of track has to be forbidden until the failure has been repaired.
A few examples of wrong-side failures modes for track circuits are shown in Table 1.

5. The simulation model

In this section a discrete-event railway simulation model is presented. The model has been developed in a C++ environment using an object-oriented programming technique. The discrete-event paradigm implies that the state of the system changes only at discrete points in time whenever an event occurs. The simulation time runs according to the sequence of scheduled events

stored in a proper calendar which is the frame of the system operation. An event is a train which enters a block section at a certain time. Train movements through the network are leaded by the chronological sequence of events generated during the simulation period according to the operational schedule, each one being consequence and cause of others. Furthermore, the model implements a set of subroutines reproducing the signalling and interlocking system and the dispatching logic which govern train movements. The model is synchronous, namely within the simulation period all trains are simulated simultaneously in the same chronological order as in reality.
Input data on the infrastructure, timetable and trains are organised by three modules. A forth module contains the safety rules implemented through the signalling and interlocking system, upon which movement authority is issued. This module implements methods to reproduce the interaction between trains, signals and movable elements (points). Details about the different parts of the simulation model are provided in the following subsections.

5.1 *Infrastructure data*

The railroad network has been modelled as a link-oriented graph (Hansen and Patch, 2008), with each link containing all information on the infrastructure which is relevant to evaluate the system. The railroad is separated into links connected by nodes. Each link represents a block section defined as a section of track bounded by two consecutive block signals (Theeg and Vlasenko, 2009). A block section is characterised by its length and the maximum permitted speed. Furthermore the initial and ending nodes specify the normal direction of travel. For each link is also specified whether a set of points is provided to allow trains changing track. Finally the system state is associated with the state of each block section which could be either occupied (or out of use), or clear. A node could represent either a signal or a station, and contains only information about its location in the network.
The main attributes which characterise nodes and links are detailed in Table 2 and 3 respectively.

Nodes	
ID_Node	Identifier of the node
Mileage	Location of the node in the network
Node_Type	Signal/Station

Table 2. Nodes' attributes.

Block Sections	
ID_Block Section	Identifier of the block section
ID_Node_i	Identifier of the initial node
ID_Node_j	Identifier of the ending node
Speed	Maximum permitted speed
Length	Block section length

Switch	Integer variable indicating whether the block section contains a point or not
State	Integer variable indicating whether the block section is occupied or clear

Table 3. Block sections' main attributes.

Each block section is considered as a unique resource with capacity one since it can be occupied by only one train at time. This safety condition is guaranteed by the principle of train separation implemented through the signalling system as described in the next section.

5.2 *Signalling and interlocking system*

This module implements a set of rules describing the behaviour of the signalling and interlocking system and its interactions with trains. The signalling system is responsible for the safe regulation of traffic on open lines, station, junction areas, and all interlocking areas. Within the interlocking area points are used to set paths allowing trains to change track. Here a fixed block signalling system is considered, according to which tracks are divided into block sections. Since movement authority can be issued only at fixed signals, each block section can be occupied by only one train at time. Furthermore, a train can enter a block section only if the block section is not occupied by another train, and only after the permission to proceed has been given through the track side signal. The signalling and safety system relies on the train detection system to issue any movement authority. The interaction between trains and signals is described by means of a set of functions to check and update block sections state according to train movements. Therefore two main functions need to be implemented into the model in order to simulate the signalling system:

- *Checking* function, to detect the state of a block section (clear or occupied)
- *Updating* function, to change the state of the block sections involved by trains movements.

Whenever a train requires a block section, the checking function is called to check the state of the system and the nature of the next event that can be scheduled is determined accordingly. In particular the principle of train separation for a fixed block system is implemented in the model in order to keep the safety distance between consecutive trains. As shown in Figure 1, train B can proceed at normal speed only if the two block sections ahead are empty. Otherwise the train will stop or slow down if the next block section, or the following one is occupied respectively.

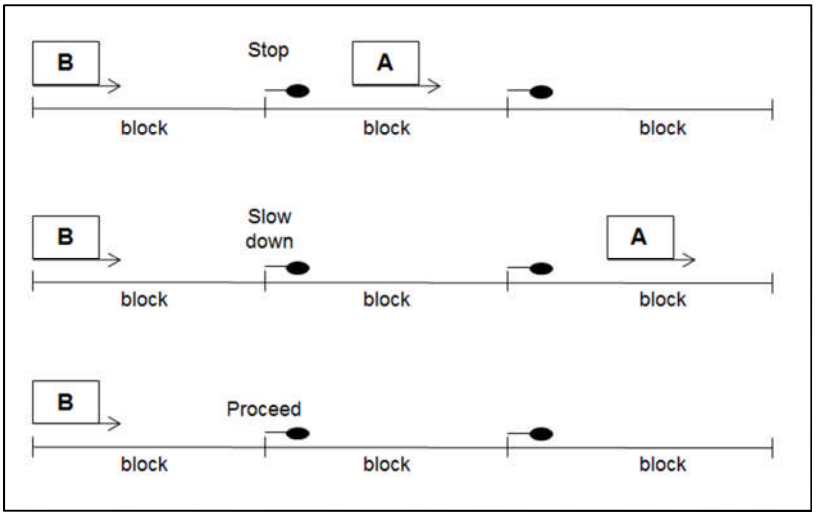


Figure 1. Principle of train separation in fixed block system.

The main elements of the interlocking system are points which can be set to normal or diverging position according to the route selected by the dispatcher. Points can connect two parallel block sections for crossing/overtaking purposes, or split a single track into one or more tracks for branching/merging different lines (junctions). When a train is approaching a block section that includes a set of points a reservation strategy is applied according to which points are set in the proper position (normal or diverging) and the state of the involved block sections is changed consequently. Points are very often a critical part of the system where conflicts between trains may arise. Therefore a dispatching logic needs to be implemented in order to solve conflicts. In particular an “event-based Boolean logic” is adopted here, which consists of a set of operational rules to be chosen according to the characteristic of the events involved. The aforementioned reservation strategy relies on the dispatching rules implemented. In the present model trains can be dispatched either on a “first-come first-served” basis, or a priority basis.

5.3 *Trains and operational data*

Trains run through the railroad according to a number of given path. A train path is defined as a sequence of block sections between the origin and destination stations. If a cyclic timetable is considered, then dwell times at stations with a scheduled stop, and headways between trains, are the input data to generate the timetable. Therefore each train is assigned to a specific path characterised by a sequence of departure and stop events.
Trains are classified as fast or slow depending on their maximum speed. The continuous motion of train movement is approximated as a discrete process, thus reducing the computational effort. Calculation of travel time is based upon the block sections length and the maximum speed allowed on the block sections. Acceleration and deceleration phases are evaluated as a percentage of the train maximum speed according to train’s acceleration and deceleration rates.

5.4 *Simulation process*

In a preliminary phase, input data on the infrastructure, trains and routes are introduced into the model so that the network layout and the operational schedule for the undisrupted scenario are defined.
The dynamic of the system is based on the sequence of events dynamically generated during the simulation according to the predefined operational schedule (paths with stops). Each train advances through the network according to its path by sequentially entering consecutive block sections and causing the system state to change. The interaction among trains, and then the order of events, is governed by the rules implemented through the signalling and interlocking system. Thus each event is consequence of a previous event and will cause another event to happen.
During each simulation events are processed in a chronological order (synchronous simulation). To keep track of the simulation time a variable t called simulation clock is used, which advances every time an event occurs. In order to define the ending condition for the simulation, a constant value P called simulation period is introduced by the user so that the simulation stops when t=P.
The events generated during the simulation are stored into a proper list called Future Events List (FEL) and processed at their scheduled time. In a disrupted scenario, conflicts may arise between trains and some events may not occur at their scheduled time. In this case unfeasible events are stored in a waiting list called QUEUE, until they can finally be processed when the safety conditions are verified. Conflicts between trains are solved according to the dispatching logic implemented. In particular a first-come first-served rule is applied here in order to dispatch trains movement. The main loop of the simulation process is summarised in Figure 2.

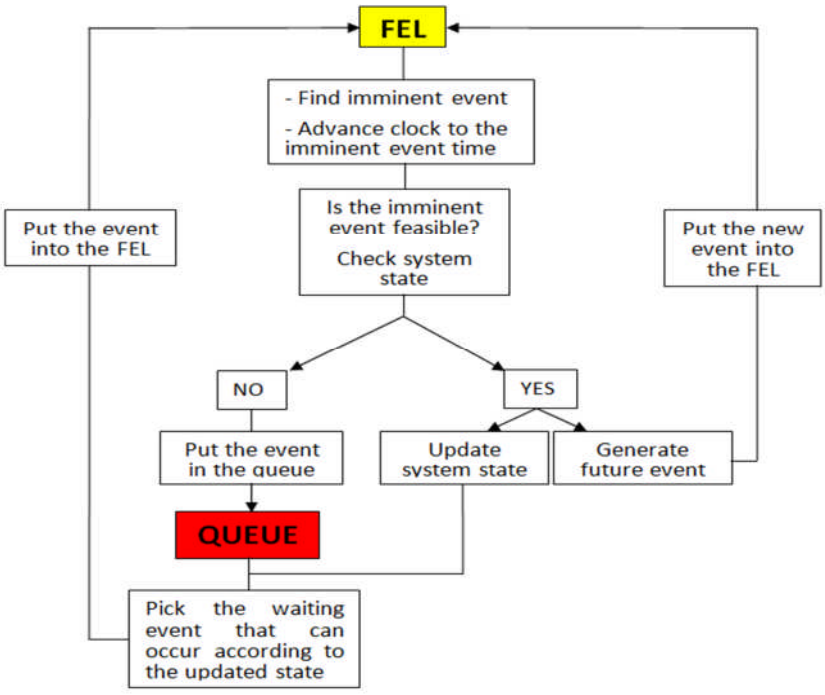


Figure 2. Main simulation loop.

The event with the lowest scheduled time is picked from the list and its feasibility is checked according to the current system state. If the event can occur at its scheduled time, then a new event involving the same train is created according to the train's path, and stored into the FEL. Once the imminent event occurs, the system state is updated accordingly. Therefore the new system state may trigger one of the waiting events stored in the QUEUE. If the imminent event is not feasible, then it is automatically stored into the waiting list and processed only when the conditions for the event to occur are verified, namely the requested resource is available. The main aim of the model is to evaluate the impact of different network layouts on system reliability, in particular different point distributions, when failures occur. During simulation failures are generated randomly from a list of failure events. For each failure a compensating action is defined, which corresponds to a particular remedial strategy implemented into the model. Therefore, according to the type of failure, one or more block sections will become unavailable or affected by speed restrictions for a certain period of time, or, if the failure severity is low no action would be required. If a failure occurs which implies a block section to be out of service, a remedial strategy is selected so that the pair of points which allow restricting the shortest section of track are set to a diverging position in order to divert traffic around the obstructed area (Figure 3).

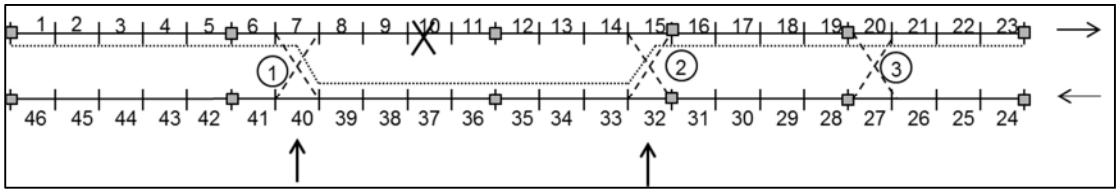


Figure 3. Selection of points and alternative path (remedial strategy).

As a consequence some of the original train paths will be no longer feasible. Thus an alternative sequence of block sections is identified according to the location and position of points. Once failures are introduced into the system, events keep being generated and processed according to the new alternative paths.
During each simulation the model keeps track of the delayed events in order to evaluate the total delay time D.
The model has been applied to simulate railway operations in a section of the UK network for a number of different scenarios. Results of the simulations are detailed in section 6.

6. Numerical example

This section provides numerical simulations for different locations of switches and disrupted scenarios. Simulations have been run for a section of the UK railway network between Harpenden and Hendon stations (Figure 4). The section is double track, one for each direction of travel; it connects six stations and covers 17,625 miles. The section has been divided into 46 links, one for

each block section, connected by 47 nodes. Node can be either signals or stations.

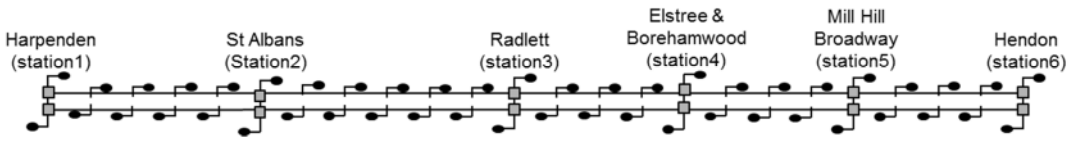


Figure 4. Network representation.

On this section, trains run according to a cyclic timetable, therefore scheduled stops repeat periodically during an operational day. Train paths are specified according to 8 different routes defined for the undisrupted scenario. The simulation period has been set to 24 hours.
The simulation model is still at an early stage, therefore it has been applied to simulate train operation in a simple network in order to be validated. A reasonable expectation is that the service reliability in terms of delay would benefit from a regular and more dense distribution of points.
The above conjecture can be initially tested for three different distributions of switches and two perturbed scenarios. In real situations, disruptions can occur every time and everywhere in the network, in particular technical failures may affect any block section at any time with a probability that depends on the failure rate. Therefore it is worth specifying that thousands of simulations need to be run for each infrastructure configuration in order to obtain a representative average delay and reliable information on the system behaviour. Location of switches for the considered infrastructure configurations is shown in Figure 5.

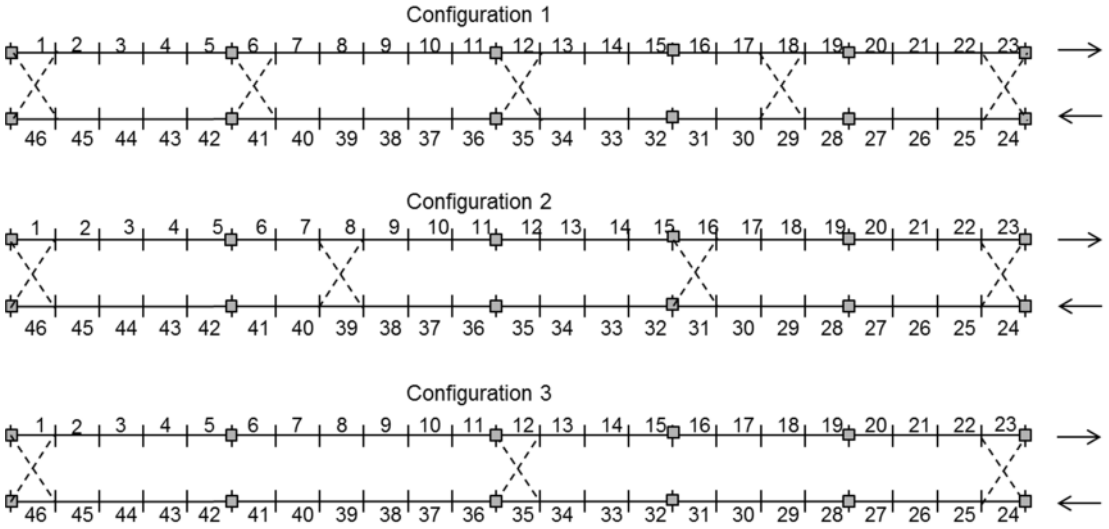


Figure 5. Infrastructure configurations.

Each configuration provides for a different number of switches: configuration 1 contains five sets of points, while configurations 2 and 3 consist of four and three sets of points, respectively. However, for all the infrastructure layouts, switches are equidistant. Furthermore two different perturbed scenarios have

been simulated, both implying the involved block sections to be out of service (Table 4).

Scenario	Identification	Block section	Time to repair	Compensating action
(a)	Track Circuit	10	1 hour	Closing section
(b)	Track Circuit	17	1 hour	Closing section

Table 4 Failure data

Simulation results for each infrastructure configuration and disrupted scenarios are compared in Table 5.

Configuration	Scenario (a)		Scenario (b)	
	Remedial strategy	Total delay time (sec)	Remedial strategy	Total delay time (sec)
(1)	Closing 7 to 11	562	Closing 13 to 17	350
(2)	Closing 9 to 15	1211	Closing 17 to 22	1937
(3)	Closing 2 to 11	1394	Closing 13 to 22	2008

Table 5. Simulation results

From a first analysis, simulation results seem to confirm that the shorter the distance between consecutive points, the smaller the delay. As a matter of fact, the infrastructure layout, which allows to overcome the obstructed area by closing a shorter section of track (configuration 1), implies a smaller delay. However it must be noticed that for the same infrastructure configuration, the delay values registered in the two perturbed scenarios are different. This confirms the need to explore a wide range of failure scenarios in order to obtain a representative average delay to be associated with each infrastructure configurations. Furthermore, in order to optimise the search of a near optimal distribution of points a structured optimisation algorithm, such as a genetic algorithm, will be implemented into the model. Obviously additional points would imply also initial and maintenance costs which must be considerate. Therefore the sub-optimal solution would be the result of a multi-objective optimisation process aiming at maximising the service reliability in a cost effective manner.

7. Conclusions

The increasing demand of railway transport in terms of both passengers and freight is a challenge for the network operators. Increasing capacity and providing high level of performance is challenging because of the trade-off between capacity and service reliability. The basic idea of this work is that

additional switches would increase network flexibility and therefore improve the ability of the system to cope with disturbances with obvious consequence on service reliability. In order to investigate the relation between service reliability and network flexibility, a discrete-event simulation model has been developed. Railway operation in a section of the UK rail network has been simulated for different infrastructure configurations under a disrupted scenario, and results are discussed.

Acknowledgments

The authors gratefully acknowledge the support of Network Rail, the Royal Academy of Engineering and The Lloyd's Register Foundation[1] (LRF).

References

1. Andrews J.D. and Moss T.R., *Reliability and Risk Assessment*, 2nd ed., Professional Engineering Publishing (2002).
2. Carey, M. and Carville, S., Testing schedule performance and reliability for train stations, *Journal of the Operational Research Society*, Vol. 51, pp. 666-682, (2000).
3. Demitz J., Hubschen C. and Albrecht C., Timetable stability – using simulation to ensure quality in a regular interval timetable. In *Computer in Railways IX*, ed. J. Allan, R. J. Hill, C. A. Brebbia, G. Sciutto and S. Sone. Southampton, United Kingdom: WIT Press (2004).
4. Dicembre A. and Ricci S., Railway traffic on high urban corridors: capacity, signalling and timetable. *Journal of Rail Transportation Planning and Management*, 1, 59-68. (2011).
5. Gill D. C. and Goodman C. J., Computer-based optimization techniques for mass transit railway signaling design. IEE Proc.-B, Vol. 139, no. 3, pp. 261-275, (1992).
6. Hansen I. A. and Patch J. (Eds), *Railway Timetable and Traffic*. Eurail press, (2008).
7. Hill R. J. and Bonf L. J., Modelling moving-block railway signaling systems using discrete-event simulation. Railroad Conference, Proceedings of the 1995 IEEE/ASME Joint, pp. 105-111, (1995).
8. Ho T.K. and Mao B.H., Reliability evaluations of railway power supplies by fault-tree analysis, *Electric Power Application, IET,* 1 [2] 161-172 (2007).
9. Luethi M., Weidmann U.A., Laube F. B., Medeossi G., Rescheduling and Train Control: A New Framework for Railroad Traffic Control in Heavily Used Networks, Proceedings of the Transportation Research Board 86th Annual Meeting, Washington DC, January 21-25, (2007).
10. Middelkoop D. and Bouwman M., Testing the stability of the rail network. In *Computer in Railways VIII*, ed. J. Allan, R. J. Hill, C. A. Brebbia, G. Sciutto and S. Sone. Southampton, United Kingdom: WIT Press, (2002).
11. Middelkoop D. and Bouwman M., SIMONE: Large Scale Train Network Simulation. In Proceeding of the 2001 Winter Simulation Conference, ed. B. A. Peters, J. S. Smith, D. J. Medeiros and M. W. Rohrer, (2001).

[1] Lloyd's Register Foundation supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

12. Panja S.C. and Ray P.K., Failure Mode and Effect Analysis of Indian Railway Signalling System, *International Journal of Performability Engineering,* 5 [2] 131-141, (2009).
13. Salido M. A., Barber F. and Ingolotti L., Robustness for a single railway line: analytical and simulation methods. *Expert System with Applications*, 39, 13305-13327, (2012).
14. Vromans M.J.C.M., Dekker R. And Kroon L.G., Reliability and heterogeneity of railway services. *European Journal of Operation Research*, 172, 647-665, (2006).

Using Deep Belief Networks for Predicting Railway Operations Failures

Olga Fink and Ulrich Weidmann
Institute for Transport Planning and Systems, ETH Zurich,
Wolfgang-Pauli-Str. 15, 8093 Zurich, Switzerland

Abstract

European passenger and freight railway demand has grown significantly during the last decade. Railway operators have increased service frequencies and network connectivity to serve this demand, but these adjustments have increased system complexity and decreased buffering capacities. As a result railway systems are now operating closer to their stability boundaries and individual failures or delays are more likely to propagate through the railway network.

A key countermeasure against delays is increasing the reliability and availability of rolling stock and infrastructure systems. Therefore, many of these systems have been equipped with advanced monitoring and diagnostic tools. These tools allow operators to efficiently identify and prevent potential failures, analyse causes of past failures and better understand the impact of failures on railway system operations.

Most diagnostic systems today provide a great deal of high dimensional and dynamic data that is non-linear and is often difficult to handle with physical or rule-based models and methods. Therefore machine learning techniques are increasingly applied to extract patterns in diagnostic data.

This paper demonstrates how deep belief networks (DBN) can be applied to predict potential operational disruptions caused by railway car door systems based on real discrete-event diagnostic data. DBN are a special type of artificial neural network that are able to recognize complex patterns and features in data. The DBN algorithm used in this research achieved a prediction accuracy of 96%. The DBN was shown to perform better than a feedforward neural network trained with genetic algorithms. In contrast to the DBN, the feedforward neural network could not discriminate between the patterns from different classes and showed a performance in the range of a random classifier.

1. Introduction

European passenger and freight railway demand has grown significantly during the last decade. Railway operators have increased service frequencies and network connectivity to serve this demand, but these adjustments have increased system complexity and decreased buffering capacities. As a result railway systems are now operating closer to their stability boundaries and individual failures or delays are more likely to propagate through the railway network.

A key countermeasure against delays is increasing the reliability and availability of rolling stock and infrastructure systems. Therefore, many of these systems have been equipped with advanced monitoring and diagnostic tools. These tools are designed to assist operating and maintenance personnel in handling faults and failures and thereby help reduce down time.

Diagnostic tools are also often used to monitor system conditions based on the principle that changes in the condition of a system can be observed in deviations of directly or indirectly measurable parameters. Several research studies focussed on developing algorithms to process system condition data and use this information to derive maintenance recommendations [1, 2].

There are two main types of diagnostic systems: continuous and event-driven. Continuous systems monitor the state of the system or process continuously and therefore provide a large amount of data, while event-driven systems only collect data at random points of time when a predefined event occurs.

Diagnostic systems generally provide huge amounts of highly dimensional, dynamic and non-linear data. These data can be difficult to interpret and handle with statistical, rule-based and physical models. It is assumed that there is a large amount of structure in the data, but the structure is too complicated to be represented by a simple model. Therefore data-based methods are increasingly being applied to extract information from the diagnostic data. Examples include artificial neural networks and support vector machines [3]. Numerous data-based methods have been applied to detect and diagnose faults, and to anticipate future system behaviour [4-7].

In the railway sector several studies have used diagnostic data to assess system condition and fault detection for turnout systems [8-10]. Other studies have considered rolling stock door systems [11, 12]. In most of these studies data from continuously measured parameters are used to extract information and partly to project the observed patterns into the future in order to anticipate impending disruption events.

Several data-based methods have been applied to predict future system behaviour including different types of neural networks [13] and support vector machines [6]. Deep belief networks (DBN) are a particular type of neural network that has proven to be a powerful tool in the field of pattern recognition [14]. They are able to capture higher order correlations and structures that are contained in the data. However they have not yet been applied for predicting disruption events or system reliability nor have they been applied in the field of diagnostics and prognostics, especially not for railway applications.

The paper demonstrates that it is possible to extract dynamic patterns of discrete events from diagnostic data using DBN and use this information to predict the occurrence of critical disruption events. The DBN algorithm was developed and tested in a case study with real data from a rolling stock door system.

The next section presents an introduction to artificial neural networks and deep belief networks in particular. Section 3 describes the data and case study. Section 4 describes the applied algorithm and the pre-processing techniques. Section 5 presents the results. Finally, Section 6 discusses the results and presents conclusions.

2. Theoretical Background

2.1 General Concepts of Artificial Neural Networks

Artificial neural networks are self-adaptive computational algorithms that show some similarities to the learning abilities of a biological neural network [3]. They are able to deduce functional behavior from input-output mappings without being presented with underlying rules or principles. The structural components of both biological and artificial neural networks are called neurons. Artificial neurons are (mostly) nonlinear processing units. Single neurons have limited functionality and learning ability. However, when neurons are connected to form more complex structures they become universal approximators [15]. The connections between the single neurons store information and are referred to as weights [3].

Artificial neural networks have been applied to a very wide variety of problems [16]. They are ideally suited to problems for which very little or no exact information is known regarding particular functional mechanisms but where a sufficient number of input-output mappings is available. These conditions make neural networks a potentially ideal method for predicting degradation processes and failure behaviour of complex systems. Given the large amount of data available from rapidly evolving railway component diagnostic systems, neural networks are a potentially excellent method for predicting the occurrence of disruption events in railway systems.

2.2 Introduction to Deep Belief Networks

Deep belief networks (DBN) are a special type of artificial neural networks. They recognise complex patterns and features in training data. DBN are composed of several layers of Restricted Boltzmann Machines (RBM) [17].

RBM are networks of symmetrically connected neuron-like units. RBMs are also referred to as stochastic neural networks [18]. Boltzmann machines consist of two layers: a visible layer and a hidden layer (Figure 1). Each unit in the visible layer is connected to all units in the hidden layer and vice versa. However, the units within one layer are not interconnected (Figure 1). Because of this limited connectivity the networks are called “restricted”. This restriction simplifies the learning process significantly [18].

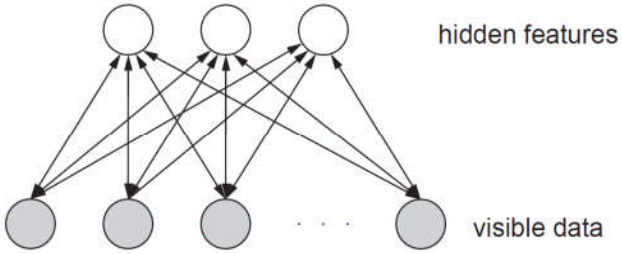


Figure 1. Network structure of a Restricted Boltzmann Machine

The visible layer contains the input parameters. The hidden layer contains the latent parameters that the networks learn from data patterns [18]. Hidden units in RBM learn the structure and features that are contained in the data. These extracted features can be, for example, parameters that influence the input data and cause a higher-order correlation between the input dimensions but cannot be directly observed or measured.

Single RBM can be joined to create more complex structures. This is the case in deep belief networks, which consist of several layers of RBM. The top two layers usually serve as associative memory of the input so that the input can be retrieved from the memory of the network [17].

The learning process for deep belief networks differs from the unsupervised learning process used for RBM. In the learning process of deep belief networks first, the network is pre-trained layer by layer in an unsupervised way. The features are learned layer by layer and the output of one layer serves as the input to the next layer [17]. After all the layers have been pre-trained, the weights between the single layers are fine-tuned in a supervised way using a backpropagation learning algorithm [17]. While backpropagation is usually not very efficient, using it in combination with the unsupervised pre-training accelerates the learning process because the weights are already pre-trained in an unsupervised way and the backpropagation is only applied for fine tuning [17].

Deep belief networks have also been applied as auto-encoders for several applications. The basic idea is to train the network to reproduce its own input. In this case the output of an intermediate layer is able to provide a lower level representation of the input data. This lower-dimensional representation can be used to perform clustering tasks on a lower dimensional level [14]. The main advantage of the auto-encoder approach is that the data does not have to be labeled to perform the classification task.

The main advantages of deep belief networks are their efficient learning algorithm, their ability to extract high dimensional features and to represent them in low dimensions, as well as their associative memory ability [17].

3. Applied data

The DBN algorithm was tested in a case study using real diagnostic discrete-event data derived from a European railway fleet. The fleet consisted of 52

train sets (the fleet was mixed between 9 coach and 11 coach train sets). All coaches had two doors on each side. The analysis period was 313 days (approximately ten months) [19]. Although the observation period was rather short, the data are considered sufficient to demonstrate the approach's feasibility given the large fleet size.

The diagnostic data was recorded automatically whenever one of the predefined events occurred. The data consisted of pertinent attributes including speed, outside temperature, overhead line voltage, etc., and depend on the affected system. A time stamp, train location (i.e. train number, car number) and actual location via GPS are also recorded for each of the occurring events [19]. Depending on the character of the occurring event, the diagnostic event belongs to one of the following categories:

- Driver action required – high priority;
- Driver action required – low priority;
- Driver information;
- Maintenance.

The events categorized as "high-priority" are immediately communicated to the driver. These are those events that can potentially result in a delay-causing event. In contrast, events that do not require an immediate action of the driver, are only provided to the maintenance crew to facilitate their maintenance activities.

Only diagnostic event data were available in the case study. Therefore it was not possible to explicitly derive the influence of maintenance actions, the age and the actual condition of the components on the occurrence of the events. Since this information was not available we assumed that these parameters and actions influence the sequence of the occurring events and the time periods between them and are thereby implicitly represented by the time series of the events themselves [19].

The operation of train doors can directly influence railway operation and schedule adherence. A door failure reduces passenger flow, causing potential train delays. Anticipating door system failures can therefore help improve reliability and efficiency.

There are 261 distinct event codes for the door system considered in this case study. These event codes indicate the specific door affected by the event. For instance, there can be four different codes for one type of event (one for each door in the car). In this research the allocation of an event to a specific door system is performed in the structure of the data and not in the coding of the events. Therefore it was possible to reduce the 261 codes to 72 distinct events. Out of the 72 events 12 require a high priority driver action [19].

Note that door system functionality can also be affected by external influences such as passengers obstructing the door. However, the functional data-driven algorithms can only predict those events originating from technical malfunctions. Therefore we have selected one of the high priority events

caused by a technical fault to demonstrate the approach feasibility in this study.

4. Applied Algorithm and Pre-Processing Techniques

4.1 Input and Output Data

Deep belief networks usually only accept binary input data [17]. However it is also possible to implement an extension to this approach and also to apply Gaussian input data. We applied Gaussian input in this research because binary representation of the data would significantly reduce the information content in the input data.

The prediction problem considered in this research was defined as a classification task. The dynamic pattern was classified as belonging to class “D” (impending operational disruption due to a failure of the door system), if within the next seven days, starting from the selected time point, at least one of the specified high priority diagnostic events having the potential to affect railway operations would occur.

If no specified high priority event occurred during the prediction period, the time pattern was classified as belonging to class “N” (no occurring events). Thus, it is a binary classification task with only two classes. While the time period of seven days might appear imprecise, it is sufficiently precise for practical purposes. This simplification increases the algorithm’s flexibility [19].

The binary classification approach increases the algorithm’s selectivity. This is a common approach in classification. Even if there are more than two classes, the algorithms often learn to distinguish one class from all other classes, which are grouped for the algorithm training process to one single class [12].

The input data patterns represent the time elapsed from the specific observation time point to the previously occurring event for each of the 72 distinct events. This approach enables us to integrate information on the time series of the occurring events into the input patterns. However, information on the density of the occurring events, which is especially important if several events occur within a short period of time, is neglected by applying this approach.

The observation time window was set at four weeks because this was considered sufficiently long to observe the occurring events and their influence on the state of the system. The information on events not observed in the current observation time window was also included because the time difference to the last occurred event is calculated.

The data-patterns were generated by moving a four-week fixed time window over the 313-day study period one day at a time. This was done to generate a sufficient number of input signals [19]. The consequence of this approach is

that the data patterns can show high similarities. These similarities are not only observed within one class, but also between the classes. This means that the classification algorithm must possess very good classification abilities to be able to discriminate between these patterns.

The input data was also normalized to have a zero mean and a unit standard deviation.

4.2 Balancing the Composition of the Data Set

Diagnostic events that can cause operational disruption are rare. Therefore the data set is highly unbalanced with many data patterns in the class “N” and comparably few in the class “D”. Algorithms trained on unbalanced data sets tend to have a weak generalization ability since they only learn to discriminate one class but not both. There are several approaches to handle unbalanced data sets [20].

In this research the data set was balanced by omitting parts of the data patterns from class “N” and including only as many data patterns from class “N” as there were from class “D” in the input data set. This approach is valid if the selected input data from class “N” sufficiently represent the distribution of the data patterns in the “N” class [19]. Since data patterns have a high degree of similarity, the assumption that the selected data patterns from class “N” are sufficiently representative for the whole class is valid. Finally, after balancing the data set, the sequence of data patterns in the data set was randomized.

4.3 Applied Algorithm

The applied deep belief network was composed of two restricted Boltzmann machines. The input with a dimension of 72 distinct input signals was presented to the first RBM in the deep belief network. Within the RBM, which itself consists of a visible and a hidden layer, the visible input is expanded to a dimension of 300 in the hidden layer. In the pre-training procedure the learning is unsupervised and takes place layer by layer for several epochs. The pre-trained output of the first RBM and its hidden layer is presented as input to the second RBM and becomes its visible layer. The visible dimension of the second layer of 300 is expanded to the dimension of 600 in the hidden layer of the second RBM (Figure 2).

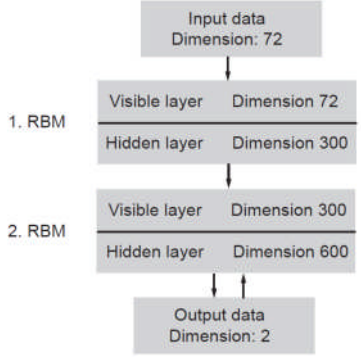


Figure 2. Structure of the applied algorithm

After the single layers have been pre-trained in an unsupervised way, the weights between the single layers are fine-tuned in a supervised way using a backpropagation learning algorithm. The error function applied within the backpropagation learning algorithm was the cross entropy (CE) error function. The CE error function performs better than the mean squared error function both in terms of computational speed and prediction performance [21]. The layers were trained for 100 epochs.

The results obtained with the DBN algorithm were compared to an alternative state of the art neural network approach. The selected network is a feedforward neural network. The learning algorithm applied in the training process is genetic algorithm (GA). Genetic algorithms are supposed to overcome limitations caused by gradient based learning algorithms such as local minima. The network was composed of two hidden layers with 40 neurons in the first hidden layer and 20 neurons in the second hidden layer. The GA was run for 20 individuals in the population over 500 generations with a crossover probability of 85%.

5. Results

The holdout technique is used to evaluate the classification performance. Holdout is a technique for assessing the generalization ability of a prediction algorithm. The holdout technique divides data into two disjoint subsets: training data and testing data. In this research we placed 90% of the data patterns in the training data set and 10% of the data patterns in the testing data set. Both subsets were assumed to be representative for the underlying data distribution of the entire data set and both subsets are independent [16].

The training data set was used to pre-train the single layers of the deep belief network and subsequently to fine-tune the weights between the layers. The testing data set was used to test how well the algorithm can generalize from data patterns that have not yet been presented to it.

The training data set consisted of 1220 data patterns and the testing data set consisted of 136 patterns.

The generalization ability was first evaluated for the general classification precision with the misclassification rate. The misclassification rate gives information on the rate of patterns that were misclassified by the algorithm, irrespective of patterns from which of the both classes were misclassified. The misclassification rate does not distinguish between the classification performance of the algorithm for different classes. In this study the misclassification rate achieved by the DBN algorithm was 3.7%. This means that 96.3% of the testing data patterns were classified correctly.

In contrast, the misclassification rate achieved by the feedforward neural network trained with GA was 43%. This performance nearly achieves the

performance of a random classifier (where the classes are assigned randomly to a data pattern). The misclassification rate was similar between the training data set and the testing data set. This means that the weak performance has not been caused by overfitting, but rather by the similarity of the data patterns and the inability of the feedforward algorithm to discriminate between the patterns from both classes.

The measures sensitivity and specificity can be used to assess the classification performance of the algorithm within the classes for a binary classification task. These measures categorize patterns of interest as “positives” and patterns from the other class as “negatives”. In this study the patterns of interest are those resulting in impending occurring events.

Sensitivity measures the algorithm’s ability to identify the positives while specificity measures the algorithm’s ability to identify the negatives [22]. Sensitivity is also referred to as the true positive rate (TRP), which is the ratio of correctly classified positive patterns to all the positive pattern in the data set (Equation 1, where TP are the true positives and P are all positives in the data set) [22].

$$TPR = \frac{TP}{P} \quad (1)$$

Specificity is also referred to as the true negative rate (TNR), which is the ratio of the correctly classified negative patterns to all the patterns in the data set (Equation 2, where TN are the true negatives and N are all negatives in the data set) [22].

$$TNR = \frac{TN}{N} \quad (1)$$

The higher the sensitivity and specificity, the better the algorithm’s performance. There is usually a trade-off between the two measures and therefore the weights between them can be adjusted depending on whether the costs of false positives (e.g. replacing parts that may not fail) are higher than those of false negatives (e.g. events were not detected by the algorithm and caused severe service disruptions). The costs for false positives and negatives are considered equally important in this study and therefore sensitivity and specificity are weighted equally [19].

The DBN algorithm achieved a sensitivity of 95.7% and a specificity of 97.0%. Although there are some marginal differences between the sensitivity and the specificity, the algorithm is not biased towards either of the two classes.

6. Conclusions and discussion

This paper describes how deep belief networks have been applied for predicting the occurrence of potential railway operation disruption events. The applicability of the approach was validated on a case study based on real discrete event diagnostic data from door systems of a railway rolling stock fleet. The prediction results obtained from the case study confirm the suitability of the proposed approach. A prediction precision of 96.3% on average was achieved.

The feedfoward neural networks trained with GA which were applied to compare the performance of DBN were not able to discriminate between the patterns from both classes and only achieved a prediction performance in the range of a random classifier. This finding confirms the good performance of DBN and their suitability to detect relevant defining features within high-dimensional data.

There are several possibilities for enhancing the DBN algorithm's performance and the significance of the results. For example, additional data such as information on the severity of the operational disruptions, the actual degradation state of the system etc., which were not available for this case study, could be integrated in the approach.

In this case study, DBN were used for a classification task with labeled data. However, it has been shown that DBN are especially powerful when labeled data is either sparse or not available at all [14]. In these cases, DBN are applied as auto-encoders. Applying DBN as auto-encoders for e.g. clustering tasks for railway operational disruptions and degradation statuses can further enhance the field of application of DBN, especially as this field of application only partially covered by other unsupervised machine learning methods.

The network structures applied in deep belief networks are usually more complex and have more layers compared to the structures applied in the current study. In this case study the deep belief networks are not used to their maximum potential since the input data are not as complex and do not have as many lower level features as in previous applications of DBN such as images [14]. Therefore, there is potential to extend the application of deep belief networks in the field of railway disruption prediction to predictions based on more complex input data.

7. Acknowledgement

The authors would like to thank ALSTOM Transportation for providing the data for this research project.

The research project is supported by the Swiss National Science Foundation (SNF) under grant number 205121_147175.

References

1. Vachtsevanos, G., *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. 2006, Hoboken, NJ: Wiley. 434.
2. Sze-jung, W., et al., *A Neural Network Integrated Decision Support System for Condition-Based Optimal Predictive Maintenance Policy.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2007. **37**(2): p. 226-236.
3. Haykin, S.S., *Neural Networks and Learning Machines*. 3rd ed. 2009, Upper Saddle River: Pearson Education. 934.

4. Shyur, H.-J., J.T. Luxhoj, and T.P. Williams, *Using Neural Networks to Predict Component Inspection Requirements for Aging Aircraft.* Computers & Industrial Engineering, 1996. **30**(2): p. 257-267.
5. Reifman, J., *Survey of Artificial Intelligence Methods for Detection and Identification of Component Faults in Nuclear Power Plants.* Journal Name: Nuclear Technology; Journal Volume: 119; Journal Issue: 1; Other Information: PBD: Jul 1997, 1997: p. Medium: X; Size: pp. 76-97.
6. Pai, P.-F., *System Reliability Forecasting by Support Vector Machines with Genetic Algorithms.* Mathematical and Computer Modelling, 2006. **43**(3-4): p. 262-274.
7. Al-Garni, A.Z. and A. Jamal, *Artificial Neural Network Application of Modeling Failure Rate for Boeing 737 Tires.* Quality and Reliability Engineering International, 2011. **27**(2): p. 209-219.
8. Ardakani, H.D., et al. *Phm for Railway System; a Case Study on the Health Assessment of the Point Machines*. in *Prognostics and Health Management (PHM), 2012 IEEE Conference on*. 2012.
9. Zhou, F.B., et al. *Remote Condition Monitoring for Railway Point Machine*. in *Railroad Conference, 2002 ASME/IEEE Joint*. 2002.
10. Garcia Marquez, F.P. and F. Schmid, *A Digital Filter-Based Approach to the Remote Condition Monitoring of Railway Turnouts.* Reliability Engineering & System Safety, 2007. **92**(6): p. 830-840.
11. Smith, A.E., D.W. Coit, and L. Yun-Chia, *Neural Network Models to Anticipate Failures of Airport Ground Transportation Vehicle Doors.* Automation Science and Engineering, IEEE Transactions on, 2010. **7**(1): p. 183-188.
12. Lehrasab, N., et al., *Industrial Fault Diagnosis: Pneumatic Train Door Case Study.* Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 2002. **216**(3): p. 175-183.
13. Lolas, S. and O.A. Olatunbosun, *Prediction of Vehicle Reliability Performance Using Artificial Neural Networks.* Expert Systems with Applications, 2008. **34**(4): p. 2360-2369.
14. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the Dimensionality of Data with Neural Networks.* Science, 2006. **313**(5786): p. 504-507.
15. Hornik, K., M. Stinchcombe, and H. White, *Multilayer Feedforward Networks Are Universal Approximators.* Neural Networks, 1989. **2**(5): p. 359-366.
16. Bishop, C.M., *Neural Networks for Pattern Recognition*. Reprinted 2005 ed. 2005, Oxford: Oxford University Press. 482.
17. Hinton, G.E., S. Osindero, and Y.-W. Teh, *A Fast Learning Algorithm for Deep Belief Nets.* Neural Computation, 2006. **18**(7): p. 1527-1554.
18. Ackley, D.H., G.E. Hinton, and T.J. Sejnowski, *A Learning Algorithm for Boltzmann Machines*.* Cognitive Science, 1985. **9**(1): p. 147-169.
19. Fink, O. and U. Weidmann. *Predicting Potential Railway Operations Disruptions Caused by Critical Component Failure Using Echo State Neural Networks and Automatically Collected Diagnostic Data*. in *92nd Annual Meeting of the Transportation Research Board*. 2013. Washington D.C. USA.
20. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*. Second ed. 2001, New York: John Wiley. 654 S.

21. Kline, D. and V. Berardi, *Revisiting Squared-Error and Cross-Entropy Functions for Training Neural Network Classifiers.* Neural Computing & Applications, 2005. **14**(4): p. 310-318.
22. Han, J., M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. 3rd ed. 2011, San Francisco, Calif.: Morgan Kaufmann. 703.

Bayesian Analysis of Electric Transmission Network Outages

Tomas Iešmantas, Robertas Alzbutas
Laboratory of Nuclear Installation Safety,
Lithuanian Energy Institute, Kaunas, Lithuania

Abstract

Electric transmission network reliability assessment is considered in this paper. The lack of coherent probabilistic treatment of network reliability is discussed and advantages of Bayesian modelling and analysis are used in order to take into account uncertainties due to differences of network parts (i.e. transmission lines located in varying environmental conditions) and resulting failure mechanisms. By analysing real North American electrical power transmission grid outage event data we show the superiority of Bayesian hierarchical models for network reliability evaluation. General results of network outage hierarchical analysis were directly transferred to evaluate reliability of specific network configuration of several lines.

1. Introduction

Although the network reliability evaluation is not something very recent, it still lacks a coherent probabilistic treatment of uncertain data and parameter estimates. Reliability or failure data of separate nodes or connecting lines are typically pooled in one sample neglecting all the variability due to differences in sources, and to worsen – raw estimates are simply plugged in to obtain the overall reliability analysis of complex network, e.g. the transmission system in a whole country [2]. Lynn et al. started to discuss it awhile ago [8], however, little efforts have been made to advance this matter.

Lynn et al. divided the state-of-the-art research of network reliability into those who compute reliability by combinatorial-like algorithms - assuming known failure probabilities but dealing with complex topologies, and those who apply statistical inference techniques to incorporate uncertainties of data sample but work with simple parallel and series configurations or some “easy” mixture of both. Extending the last case we would like to add another, more recent, trend – Complex Network Analysis, when the reliability of the grid is evaluated taking into account the topological relations between nodes and connecting lines. Nice survey of the look at power grid as a complex network can be found in a survey of Pagani and Aiello [13].

The result of such scientific community division is that analysis of complex networks is left without any proper treatment of uncertainties. Having this in mind, Lynn et al. developed a methodology of any network reliability evaluation by joining pivotal decomposition [3] together with Bayesian inference.

Following this work, reliability assessment of networks functioning in random environments started to develop, mostly due to Özekici [12, 11]. The main idea is that system functions within fluctuating weather conditions and the failure probability is conditioned on the state of the environment. It allowed incorporating uncertainties due to different states of environment. However,

the application areas were never extended to the electrical power grids where various stochastic phenomena cause random outages and cascading outages.
The purpose and contribution of this paper is the advancement of aspects stated above. We seek a coherent probabilistic treatment of grid reliability parameters by “immersing” grid failure data into Bayesian framework – its ability to naturally handle uncertainties of data, to express all information about parameter estimates and possibility to deal with more complex models led to this choice. By analysing real data of North Americas electrical power transmission grid (Section 2.1.) we go from simple analysis of lines outage intensities to intensities affected by heterogeneity of grid and surrounding environment (Section 2.2.) and apply results to estimate reliability of part of the network (Section 2.3.).
We refrain ourselves from describing the Markov Chain Monte Carlo methods which were used to implement Bayesian and hierarchical Bayesian methods employed in the analysis. Although it is interesting and at the same time not an easy task, it is out of the scope of this paper and interested reader can found necessary information in our other papers [1, 7] or textbooks of Gilks et al. [6], Ntzourfras [10] etc.

2. Bayesian analysis of transmission grid reliability data

As already mentioned in the introduction, we will analyse the electric transmission grid reliability when taking into consideration inhomogeneity in network outage statistical data. We will talk about how uncertainties in the data and estimates of reliability of separate transmission lines can be propagated through the Bayesian inference framework to obtain reliability of the some part (or configuration) of the transmission network.

2.1 Transmission grid outage data description

The data that is analysed in subsequent sections were obtained from Bonneville Power Administration database [4] that contains information about the outages, timings and causes of them. The electric transmission grid spans over the areas of California, Oregon, Idaho, Montana and Washington. Hence, the variability of environmental conditions varies significantly.
Since our research does not seek of full investigation of this particular power grid, we have confined ourselves with transmission grid of 500 kV lines. Time span of outage events is 11 years and involves 3179 events (non-planned outages) produced by 97 transmission lines (distribution of frequencies of outages for each line is presented in Figure 1).
At this stage of research we have discarded the causes of the outages in order to lay grounds for more simplistic Bayesian treatment of the outage phenomena. Further model developments might include causes through e.g. regression part. Initially, we will model these events as coming from Poisson distribution.

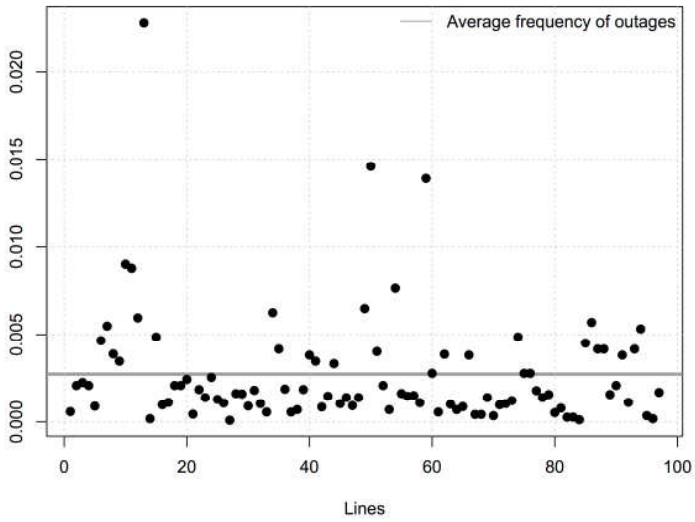


Figure 1. Outage frequencies for each line compared to average frequency

2.2 Bayesian analysis of line outage statistics

The number of outages can be modelled by the Poisson distribution. In this section we will perform analysis and validation of the various modifications of simple Poissonian model.

Suppose that the data is described by the triplet $\left(X_i, L_i, t_i\right)_{i=\overline{1,N}}$, where X_i is the number of outages collected over the period t_i for the i[th] line with length L_i (divided by 1000 km) when the number of lines is N. According to these notations, the model can be expressed as follows:

$$X_i \mid \lambda, t_i, L_i \sim Poisson\left(t_i L_i \lambda\right),\ i = \overline{1,N}; \tag{1}$$

where λ denotes the expected number of outages over one year for line of 1000 km length, hence the term $t_i L_i \lambda$ is the expected number of outages for i[th] line over time t_i when the length is L_i. In order to have a full Bayesian model, we chose so called improper prior distribution for intensity parameter:

$$\pi\left(\lambda\right) \propto 1_{(0,+\infty)}. \tag{2}$$

Resulting posterior distribution is a gamma distribution:

$$\lambda \mid X \sim Gamma\left(1 + \sum X_i, \sum L_i t_i\right); \tag{3}$$

This distribution summarizes all the information about parameter λ necessary to make inference. Expected value and standard deviation are 0.00215 and $4.73e-05$ accordingly. No normality assumptions are required (as in maximum likelihood estimation) in order to obtain 95% Bayesian credibility interval, which in this case is

$$Pr\left[0.00205 \le \lambda \le 0.00224\right] = 0.95. \quad (4)$$

This model, although is very simple and straightforward to analyse, assumes that all lines produces outages with the same intensity λ. But this assumption could be misleading, since the transmission grid is established over wide geographical area and different parts of it experience different weather loads (leads to different degree of lines deterioration), different power loads (heavier loading exposes hidden failures [9]), etc. Due to inability to account for heterogeneity in intensity of outages, simple Poisson model fails to incorporate all uncertainty properly. Hence, related risk or reliability analysis using this intensity will produce inadequate measures.

In order to account for source-to-source (or line-to-line) variability, we use hierarchical Bayesian model, which enables borrowing the strength over all samples and at the same time models each outage data sample separately. This type of model can be thought about as intermediate case between complete data pooling and no pooling. Graphically simple hierarchical structure can be represented as in Figure 2: arrows pointing downwards represent influence relations, i.e. each upper level outcome drives the process of lower level, while the information of data flow upwards and deteriorates more and more at each level.

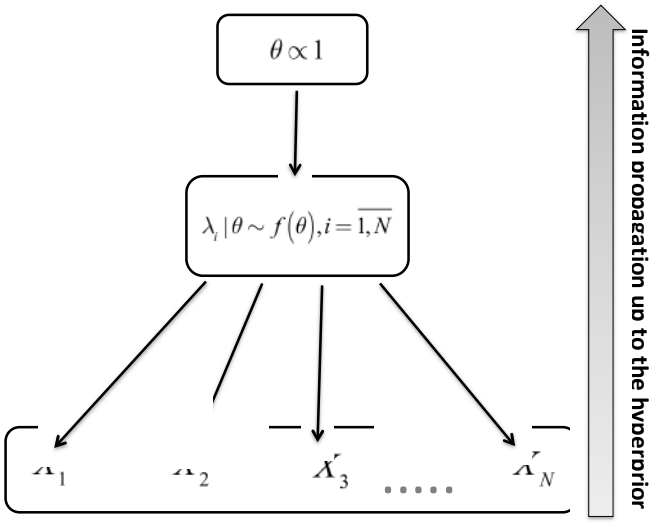


Figure 2. Graphical representation of hierarchical model and flow of info

We obtain Bayesian hierarchical model for Poissonian data as follows:

$$\begin{aligned} &X_i | \lambda_i \sim Poisson\left(t_i L_i \lambda_i\right), i = \overline{1,N}; \\ &\lambda_i | \theta \sim f\left(\lambda_i; \theta\right); \\ &\pi\left(\theta\right) \propto 1. \end{aligned} \quad (5)$$

By stating this model, we made several assumptions. First one is that distribution of the unobservable population $\left\{\lambda_i\right\}_{i=\overline{1,N}}$ is Gaussian. Second assumption is that uniform hyper-prior distribution is used for mean and

variance parameters. We might equally assume for transformed parameters $exp(\lambda_i)$ the lognormal distribution or gamma or other distribution defined on positive real axis.

We will not go into extensive mathematical formulation of the implementation of the MCMC method, however it would be useful to give a general scheme of MCMC implemented in case of, for example, lognormal distribution with parameters μ and σ.

The posterior distribution in this case could be expressed (up to a constant) as follows:

$$\pi(\lambda,\mu,\sigma \mid X) \propto \left[\prod_{i=1}^{N} e^{t_i L_i \lambda_i} \lambda_i^{X_i}\right]\left[\prod_{i=1}^{N} \frac{1}{\lambda_i \sigma} e^{-\frac{(\log \lambda_i - \mu)^2}{2\sigma^2}}\right]\frac{1}{\sigma^2}. \quad (6)$$

For this expression is hard to obtain random samples, however conditional distributions for μ and $1/\sigma$ have analytical expressions leading to easier posterior sampling. In this particular lognormal distribution case sampling scheme could be visualized as in Figure 3.

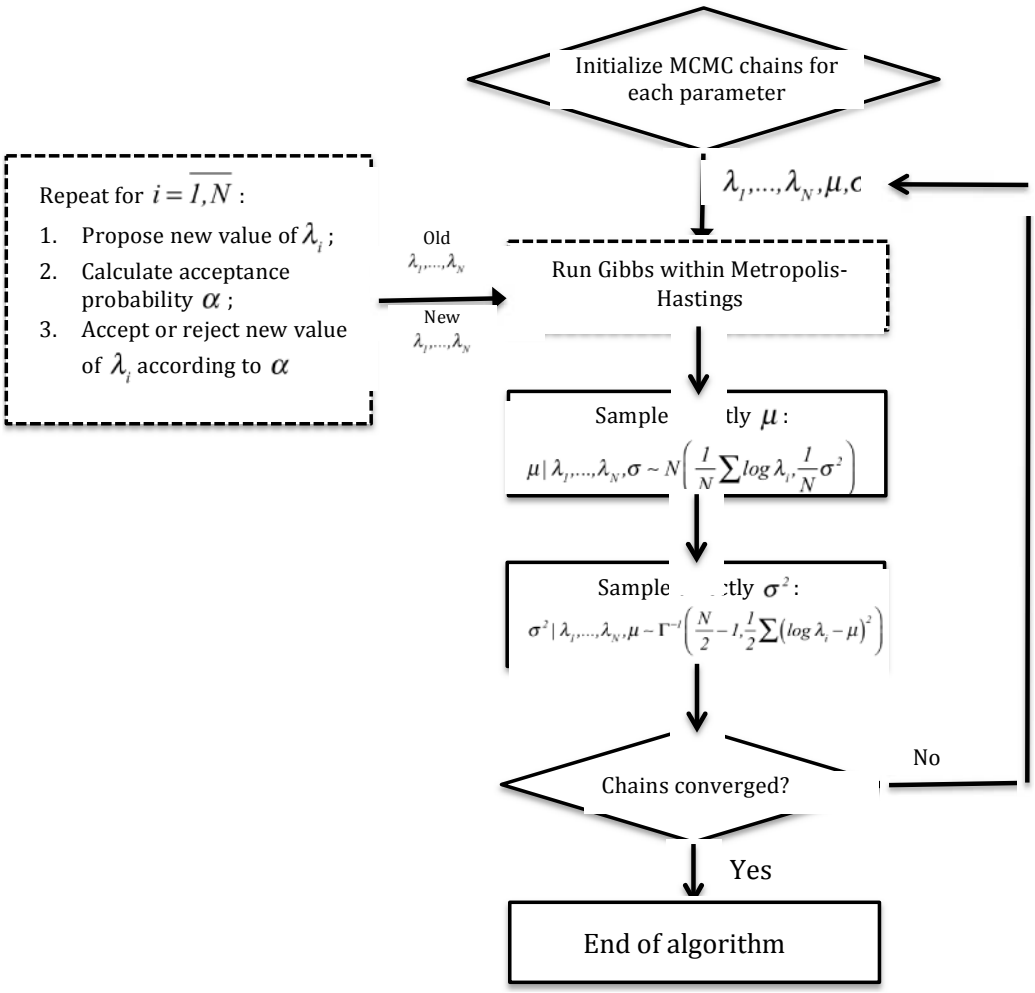


Figure 3. General posterior sampling scheme for lognormal distribution case

To select and to compare to the homogeneous Poisson distribution, Deviance Information Criterion (DIC, [14]) was employed. DIC is defined as follows:

$$DIC = -2E_{\Theta|X}\left[\log L\left(X|\Theta\right)\right] + p_D; \quad (7)$$

where Θ denotes all model parameters and p_D is a so called effective number of parameters, which is typically lower than the nominal number of parameters due to borrowing of strength under the hyperdensity [5] and can be regarded as a measure of model complexity. DIC is a measure of fitness of predictive distribution, and since the prediction of uncertainty is one of the main goals for practitioners, it is an appealing model characteristic for our assessment purposes.

Table summarises the results of fitness of different probability models for outage statistics. The smallest DIC is of hierarchical model under lognormal population distribution. It also has the smallest p_D value of all hierarchical models considered.

Distribution	Nonhierarchical	Normal	Gamma	Lognormal	Weibull
DIC	29497.24	28219.73	27158.41	26986	27160.01
pD	1.01	97.21	89.53	82	89.37

Table 1. DIC and pD values for all considered models

Replication of outages from non-hierarchical and from hierarchical (with lognormal population distribution) models confirms the superiority of hierarchical structure: mean squared error between replicated and observed data is 201 for hierarchical model and 55850 for simple. It properly incorporates within- and between- source uncertainties and, as a consequence, has better prediction properties – in case of individual as well as overall rate of outages.

Fitness of hierarchical structure implies that intensity by which outages occur varies across different transmission lines. Some lines are more vulnerable than others and this gives the information which parts of the grid should be targeted when maintenance plan is being set. Hierarchical model, as presented below, does not account for causes of outages – it shows that there is significant variability or difference between separate transmission lines. However, it can be easily extended to include information about different causes of outages by including e.g. a regression part into Poisson intensity.

The difference of two models analysed above can be clearly seen in long-term outage number predictions: simple Poisson model favours smaller expected number of outages over all (500 kV) transmission grid: for instance, the difference of predicted number of outages for 3 years is almost 300 outages and as the prediction interval increases this difference increase as well. This implies that the maintenance planning as well as prediction of loss of power load probability will be underestimated when simple Poisson model is employed.

2.3 Bayesian network reliability

This section is devoted for the demonstration of the reliability analysis of a network in terms of Bayesian framework. We will use the results of previous section: hierarchical Poisson model of electric transmission lines together with posterior distributions of outage intensities. In addition there will be used several assumptions:

- Transmission line outages are independent given their failure rates;
- Transmission lines once are failed, stay in this outage state for some time period $\left[0;T\right]$;
- No failure occur in network nodes, i.e. nodes are perfect in terms of reliability;
- Transmission line can be in two states: not failed (state 1) and failed (state 0);

It is hard to tell at what degree the first assumption holds, but in general it is not true – e.g. in cascading failures, outages are at least dependent on the previous cascade stage. However, due to lack of literature addressing these dependency issues we refrain ourselves from assuming any dependency between lines.

In addition, second assumption is not realistic for long time periods – all transmission lines sooner or latter are repaired. However, if we consider short enough time period, then the network can be treated as non-repairable system. It is generally agreed that failures separated by more than and one-hour belongs to different cascades. Hence, one-hour window could be thought as a short enough so that lines can be regarded as non-repairable components.

Having this in mind, we consider the probability that part (Figure 4) of the entire network is connected in terms of ability to deliver electricity between Monroe and Keeler in time period $\left[0;T\right]$.

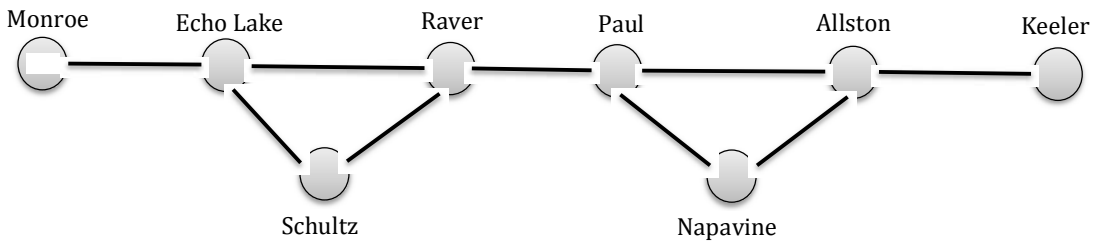


Figure 4. Structural representation of network over 8 nodes

If we denote the state of k[th] transmission line by $Y_k\left(t\right)$, the structure function $\phi\left(Y_1\left(t\right),...,Y_8\left(t\right)\right)$ represents the state of the network and posterior probability of failure is [8]:

$$F\left(t\right)=E\left[\phi\left(Y_1\left(t\right),...,Y_8\left(t\right)\right)\mid X\right]= \int\phi\left(p_1\left(t\right),...,p_8\left(t\right)\right)\pi\left(p_1\left(t\right),...,p_8\left(t\right)\right)dp_1\left(t\right),...,dp_8\left(t\right) \tag{8}$$

where $p_k(t) = Pr\left[Y_k(t) = 1\right]$. Since we are assuming hierarchical Poisson distribution over outage as statistic, we have that $p_k(t) = e^{-tL_k\lambda_k}$.
Since from the previous section we already have samples from posterior distributions of Poisson intensity parameters, it is straightforward to obtain the posterior distribution of failure of the network by feeding structural function with $p_k^i(t) = e^{-tL_k\lambda_k^i}$, where λ_k^i are random draws from posterior λ_k distribution.
In our network case we have a distribution of network failure probability expressed as a histogram in Figure 5. Comparison with maximum likelihood (ML) estimate of Poisson intensity parameter shows the impact of the hierarchical model structure – expected value of probability for the network to fail (within one hour time window) is less than the estimate obtained by ML method. By including additional variability we are able to evaluate failure without overestimation.

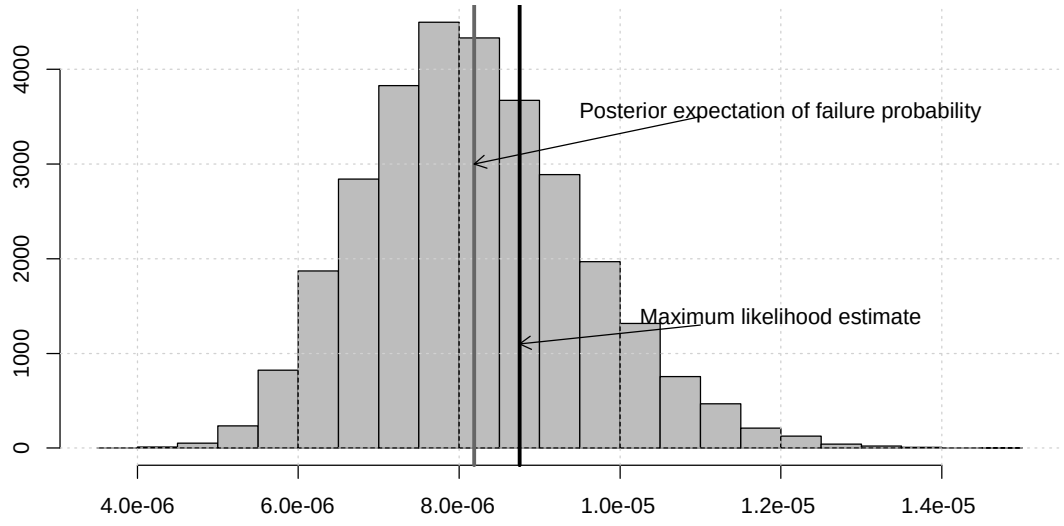


Figure 5. Posterior network failure probability distribution

3. Conclusions and further discussions

A thorough analysis of electrical power transmission grid reliability data was presented in this paper. Bayesian methods, and, in particular, hierarchical methods served as a basis all analysis. We showed that employing Bayesian methods could enhance transmission grid reliability analysis – this is due to ability to deal with uncertainties in data and parameters as well as to tackle complex hierarchical structures imposed on that data. Bayesian machinery in transmission line outage case permitted to obtain evidence about the dissimilarities of outage processes generated by transmission lines.
General results of hierarchical outage analysis were directly transferred to evaluate reliability of specific configuration of several lines – assuming we can find all the minimal cut-sets of the network it is straightforward to obtain reliability of any complex configuration. In addition, this shows, how easily

random samples from posterior distributions can be reused in further analysis of the same object.

Although analysis was performed just for North Americas' power grid, and general conclusion about hierarchical nature of outages for other countries cannot be drawn, results provide us with evidence of the presents of the phenomena. These evidences should be though as an alert for other researchers tackling the reliability issues of electric power grid, or any other complex grid.

Although we have not considered in this paper time dependency in outage events, hierarchical Bayesian methods can be easily extended to do so. Non-homogeneous Poisson or even count data time series can be brought into a picture of electric power grid reliability without difficulty. In addition, this would be translated into a time-dependent failure probability of part of the network. Such flexibility opens vast possibilities for further development of network reliability theory with a proper uncertainty handling. Hence, this means a much more accurate predictions and reliability-based maintenance planning.

Acknowledgement

This research was partially funded by the grant (No. ATE-04/2012) from the Lithuanian Research Council.

References

1. Alzbutas R. and Iešmantas T. Application of Bayesian Methods for Age-Dependent Reliability Analysis. *Quality and Reliability International*. DOI: 10.1002/qre.1482 (2013).
2. Augutis J., Krikštolaitis R., Alzbutas R., Matuzas V., Ušpuras E. Reliability analysis of the electricity transmission system in Lithuania, *Risk analysis IV*, Fourth international conference on computer simulation in risk analysis and hazard mitigation. ISSN 1470-6326, ISBN 1-85312-736-1. WIT Press, Southampton, Boston, 573-580 (2004).
3. Barlow R. E., Proschan F. *Statistical theory of reliability and life testing*. Holt, Rinehart, and Winston, New York, (1975).
4. Bonneville Power Administration Transmission Services Operations & Reliability website: http://transmission.bpa.gov/Business/Operations/Outages.
5. Congdon P. *Applied Bayesian hierarchical methods*. Chapman and Hall/CRC (2010).
6. Gilks WR, Richardson S, Spiegelhalter DJ, *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC (1996).
7. Iešmantas T. and Alzbutas R. Age-dependent hierarchical Bayesian modelling for reliability assessment under small data sample, Proc. of 11th international probabilistic safety assessment and management conference and the annual European safety and reliability conference (PSAM11 ESREL2012), Helsinki Finland, June 25-29, 2012. IAPSAM & ESRA. ISBN 978-1-62276-436-5, 2527-2537 (2012).
8. Lynn N., Singpurwalla N. and Smith A. Bayesian assessment of network reliability. *SIAM review* 40, 202-227 (1998).
9. Mili L. and Qiu Q. Risk assessment of catastrophic failures in electric power systems. *Int. J. Critic. Infrastruct*. **1**, 38-63 (2004).

10. Ntzoufras I, *Bayesian Modelling Using WinBUGS*. Wiley (2009).
11. Özekici S., Soyer R. Bayesian analysis of Markov Modulated Bernoulli Processes. *Math Meth. Op. Res.* **57**, 125-140 (2003).
12. Özekici S., Soyer R. Network reliability assessment in a random environment. *Nav. R. Log*. **50**, 574-591 (2003).
13. Pagani G.A., Aiell, M. *The power grid as a complex network: a survey.* Tech. Rep. ArXiv preprint arXiv:1105.3338 (2011).
14. Spiegelhalter D.J., Best N.G., Carlin B.P. and van der Linde A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* **64**, 583-639 (2002).

Predictive and diagnostic analysis of an holdup tank by means of Dynamic Bayesian Networks

Daniele Codetta-Raiteri[1], **Luigi Portinale**[1]
[1]DiSIT, Computer Science Institute, University of Piemonte Orientale, Alessandria, Italy

Abstract

In dynamic reliability evaluation, the complete behaviour of the system has to be taken into account. In this paper, a benchmark taken from the literature is examined. To this aim, we exploit Dynamic Bayesian Networks (DBN) extending standard Bayesian networks by introducing a discrete temporal dimension. The goals are the prediction of the system unreliability and the computation of diagnostic indices. Because of the achievement of such goals, we propose DBN to be a valid approach for dynamic reliability evaluation.

1. Introduction

We talk about dynamic reliability [1] when the system configuration changes during the mission time. In these cases, we may have to consider the whole system behaviour in order to evaluate the reliability. This means modelling the normal functioning of the system, the occurrence of component failure events and their effect on the system functioning. Combinatorial models [2] such as *Fault Trees* and *Reliability Block Diagrams* can only represent combinations of component failure events assumed to be independent. The complete system behaviour can be represented by means of state space based models [2], such as *Markov Chains* or *Petri Nets*. They rely on the specification of the whole set of the possible system states, so that the stochastic behaviour of each component may depend on the state of all the other components. However, their use in dynamic reliability may determine the state space explosion making the model analysis unpractical because of the high computing cost (and time).

Bayesian Networks (BN) [3] are an interesting trade-off between combinatorial and state space based models; in particular, *Dynamic Bayesian Networks* (DBN) [4] provide an explicit discrete temporal dimension: a DBN represents the system at several discrete time slices, and conditional dependencies among variables at different slices are introduced to capture the temporal evolution. Both BN and DBN have been recently investigated as very promising formalisms for dependability and reliability analysis [3, 5, 6]. We argue that DBN are a possible and suitable approach to examine dynamic reliability cases; we show this point by investigating the analysis of a specific benchmark taken from the literature [1]. The benchmark is a system consisting of a tank containing some liquid whose level is influenced by a controller commanding two pumps and one valve, with the aim of avoiding the dry out or overflow of the liquid. In the past, the benchmark was evaluated by means of Monte Carlo simulation [1] and Petri Nets [7, 8]. In this paper, the system is modelled as a DBN, with the purpose of

computing the system unreliability (the original goal of the benchmark [1]), and diagnostic indices which are an additional possibility offered by DBN.

The paper is organized as follows: Section 2 contains the related work about this benchmark; Section 3 describes the system behaviour; Section 4 provides the essential notions about the DBN formalism; Section 5 describes the DBN model of the system; finally, Section 6 reports the results of the model analysis.

2. Related work

The benchmark is specified in [1], where the system unreliability is evaluated by means of the Monte Carlo simulation. In [7] the benchmark is modelled as a *Generalized Stochastic Petri Net* (GSPN) [2]. The GSPN model can undergo analysis, but it suffers from two approximations: **1)** the liquid level is discretized into few intermediate levels, because only discrete variables can be represented as the number of tokens (marking) inside places; **2)** some deterministic timed events such as the action of the pumps on the liquid level, are considered as stochastic events. Still in [7], the benchmark is modelled and simulated as a *Fluid Stochastic Petri Net* (FSPN) [9] including also fluid places which directly represent continuous variables such as the liquid level in the tank. Finally, in [8], *Stochastic Activity Networks* (SAN) [10] are applied in order to model and simulate the benchmark. SAN extend Petri Nets introducing input or output gates able to express complex conditions and effects about the firing of transitions, compacting the model as a consequence. SAN can represent float variables by means of extended places.

In [1], other versions of the benchmark are presented and are characterized by particular features (state dependent failure rates, failure on demand, repair). They are evaluated using Petri Net based approaches in [7, 8], and using DBN in [11].

3. The case study

The system (Figure 1.a) is composed by a tank containing liquid, two pumps (P1, P2) to fill the tank, one valve (V) to remove liquid, and the controller (C) monitoring the liquid level (H) and switching P1, P2, V on or off. The state of P1, P2, V can be ON, OFF, Stuck ON (S_ON), or Stuck OFF (S_OFF). Initially H is equal to 0, with P1 and V in state ON, and P2 in state OFF; since both pumps and the valve have the same level variation rate (Q=0.6 m/h), H does not change while the initial configuration holds (Tab. 1). The cause of a variation of H may be the occurrence of a component failure during the ON or OFF state. The failure probability obeys the negative exponential distribution: the failure rate λ of P1, P2 and V is equal to 0.004566 h^{-1}, 0.005714 h^{-1} and 0.003125 h^{-1}, respectively. The effect of the failure is the stuck condition, while the state transitions toward S_ON and S_OFF, are uniformly distributed (Figure 1.b).

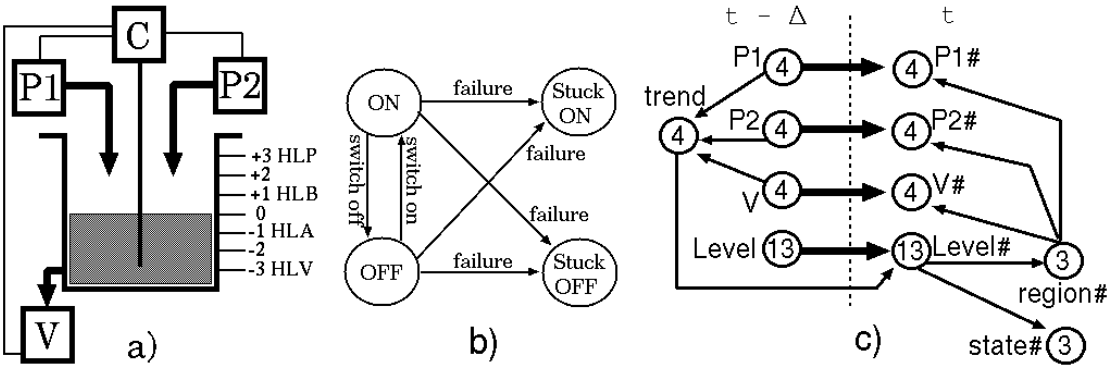


Figure 1: **a)** The system scheme. **b)** The possible states of P1, P2 and V. **c)** The graph of the DBN model.

Comp. states				Comp. states				Comp. states			
P1	P2	V	effect	P1	P2	V	effect	P1	P2	V	effect
OFF	OFF	OFF	=	OFF	ON	ON	=	ON	ON	OFF	↑↑
OFF	OFF	ON	↓	ON	OFF	OFF	↑	ON	ON	ON	↑
OFF	ON	OFF	↑	ON	OFF	ON	=				

Table 1: The effect on H in each state configuration (S_ON and S_OFF have the same effect of ON and OFF respectively).

Tab. 1 shows how H changes with respect to the current configuration of the component states; C believes that the system is correctly functioning while H is inside the region between the levels denoted by HLA (-1 m) and HLB (+1 m) shown in Figure 1.a. If H reaches HLA, then C orders to P1 and P2 to switch on, and V to switch off (**order n. 1**), with the aim of increasing H and avoiding the **dry out**; this event occurs when H reaches the level denoted as HLV (-3 m). If a component is stuck, it does not obey the controller order and maintains its current state. The other system failure condition is the **overflow**; this happens when H reaches HLP (+3 m). If H reaches HLB, C orders to P1 and P2 to switch off, and V to switch on (**order n. 2**), with the aim of decreasing H and avoiding the overflow.

4. Basic notions about DBN

BN are defined by a directed acyclic graph (DAG) in which nodes correspond to discrete random variables having a conditional dependence on the parent nodes. DBN extend BN by providing a discrete temporal dimension. The advantage with respect to a classical probabilistic temporal model like *Markov Chains*, is that a DBN is a stochastic transition model factored over a number of random variables. While a DBN can in general represent semi-Markovian stochastic processes of order $k-1$, providing the modelling for k time slices, the term DBN is usually adopted when $k=2$. If so, the Markovian assumption holds and only 2 time slices are considered in order to model the system tem-

poral evolution: the slice at time t depends only on the previous slice at $t - \Delta$, and is conditionally independent of the past ones (Δ is the time discretization step). An example of DBN is shown in Figure 1.c.

In a DBN, we can distinguish between two kinds of arcs: intra-slice and inter-slice arcs establishing dependencies between variables in the same time slice, and dependencies between variables in different time slices, respectively. The variables characterized by a temporal evolution, have two instances, one for each time slice. connected by a temporal arc graphically appearing as a thick line. Inter-slice arcs connecting two instances of a variable are called *temporal arcs*. For instance, in Figure 1.c, $P1$ and $P1\#$ are present in the time slices $t - \Delta$ and t respectively, and are connected by the temporal arc $(P1, P1\#)$. This means that P1 may change its state during the interval Δ between two consecutive time steps.

The dependencies of a certain DBN node are quantified in terms of conditional probabilities and are stored in its *Conditional Probability Table* (CPT). The probability in every CPT entry has to be set according to the state of the parent nodes (possibly including the other instance of the node).

Analysis. Let X^t be a set of variables at time t and $y_{a:b}$ any stream of observations between the time points a and b (a set of instantiated variables Y_i^j with $a \leq j \leq b$). The following tasks can be performed over a DBN:
• *Prediction*: computing $P(X^{t+h}|y_{a:b})$ for some horizon $h > 0$, i.e. predicting a future state taking into consideration the observation up to now (if $h = 0$ the task is more properly called **Filtering** or **Monitoring**);
• *Smoothing*: computing $P(X^{t-l}|y_{a:b})$ for some $l < t$, i.e. estimating what happened l steps in the past, given all the evidence (observations) up to now.

In this work, the DBN model is designed and analyzed by means of the RADYBAN software tool [12]. In particular, for the analysis, we resort to the *Junction Tree* (JT) algorithm based on the construction of a classical BN inference data structure called junction or join tree [4]. The JT algorithm returns exact results for both the above tasks.

5. Modelling the benchmark

State of components. The DBN model of the benchmark is shown in Figure 1.c. The state of P1 is represented by the variables $P1$ and $P1\#$. Their value can be 0, 1, 2, or 3, in order to represent the states OFF, ON, S_OFF, S_ON, respectively. The variable $P1$ in the time slice $t - \Delta$ does not depend on any other variable ($P1$ is a root node). Therefore its CPT (Section 4) simply provides the initial probability distribution of the four possible values. In particular, the value 1 has probability 1 in order to express that the initial state of P1 is ON. The variable $P1\#$ in the time slice t depends on $P1$ and $region\#$ in order to model that the current state depends on the state in the previous time step, and on the current region determining the command currently provided by C

(Section 3). So, the variable $region\#$ is ternary: the value 0 corresponds to the order n. 1, 1 represents the absence of orders (correct region), 2 corresponds to the order n. 2 (Section 3). The CPT of $P1\#$ (Tab. 2) contains the probability distribution of the possible values of $P1\#$ given all the possible combinations of the values of $P1$ and $region\#$. Let us consider the entries of this CPT:
• the **entry n. 1** is the case where $P1 = 0$ and $region\# = 0$; this means that P1 was OFF in the previous time step and C orders the pumps to switch on in the current time. So, the probability that $P1\# = 0$ (P1 is currently OFF) is null because P1 can not ignore the order if it is not stuck. The probability that $P1\# = 1$ (P1 is currently ON) is the probability that P1 does not fail (reliability of P1) during the transition between the time slices $t - \Delta$ and t, according to the negative exponential distribution, the rate λ (Section 3) and the time step Δ (the value of Δ will be specified in the following). The probability that $P1\# = 2$ (P1 is currently S_OFF) is half of the probability of failure, because the probability to turn S_ON or S_OFF after the failure, is uniformly distributed (Section 3). The probability that $P1\# = 3$ (P1 is currently S_ON) is computed in the same way. The sum of the probabilities in the entry n. 1 and in the following entries has to be 1. For the sake of brevity, the lines characterized by null probability are omitted in the CPTs.
• The **entry n. 2** is the case where P1 was OFF and C provides no order. Therefore $Pr\{P1\# = 1\}$ is null, while $Pr\{P1\# = 0\}$ is the reliability of P1 during the time step Δ. $Pr\{P1\# = 2\}$ and $Pr\{P1\# = 3\}$ are computed in the same way as in the entry n. 1 and the following ones, up to entry n. 6.
• The **entry n. 3** is the situation where P1 was OFF and C orders the pumps to switch off. So, $Pr\{P1\# = 0\}$ is the reliability of P1 during the time step Δ, while $Pr\{P1\# = 1\}$ is null because of the order from C.
• In the **entry n. 4**, P1 was ON and the command is to turn ON. Therefore $Pr\{P1\# = 0\}$ is null, while $Pr\{P1\# = 1\}$ is the probability that P1 does not fail (reliability).
• In the **entry n. 5**, P1 was ON and no commands are provided, so the same probability distribution as in the entry n. 4, holds.
• The **entry n. 6** is the case where P1 was ON and C orders the pumps to switch off. Therefore $Pr\{P1\# = 0\}$ is the component reliability, while $Pr\{P1\# = 1\}$ is equal to 0.
• In the **entries n. 7, 8, 9**, P1 was in the S_OFF state ($P1 = 2$) in the previous time step. Since P1 is not repairable, P1 maintains such state in the current time step, ignoring any command from C. Therefore $Pr\{P1\# = 2\}$ is equal to 1 in all the entries, while the probabilities of the other values are null.
• In the **entries n. 10, 11, 12**, P1 was S_ON ($P1 = 3$), so $Pr\{P1\# = 3\}$ is equal to 1 in all such entries.

The states of P2 are modelled in the same way by the variable $P2\#$ depending on $P2$ and $region\#$. The states of V are represented by the variable $V\#$ influenced by V and $region\#$. The CPT of $V\#$ takes into account the opposite reactions of V to the orders, and the failure rate λ of V (Section 3).

Variations to H. The variable $trend$ depends on the variables $P1$, $P2$ and V, and its value can vary between 0 and 3. The role of this variable is to represent

n.	$P1$	$region\#$	***P1#***	prob.	n.	$P1$	$region\#$	***P1#***	prob.
	0	0	1	$e^{\lambda\Delta}$		1	1	1	$e^{-\lambda\Delta}$
1	0	0	2	$(1-e^{-\lambda\Delta})/2$	5	1	1	2	$(1-e^{-\lambda\Delta})/2$
	0	0	3	$(1-e^{-\lambda\Delta})/2$		1	1	3	$(1-e^{-\lambda\Delta})/2$
	0	1	0	$e^{\lambda\Delta}$		1	2	0	$e^{-\lambda\Delta}$
2	0	1	2	$(1-e^{-\lambda\Delta})/2$	6	1	2	2	$(1-e^{-\lambda\Delta})/2$
	0	1	3	$(1-e^{-\lambda\Delta})/2$		1	2	3	$(1-e^{-\lambda\Delta})/2$
	0	2	0	$e^{\lambda\Delta}$	7	2	0	2	1
3	0	2	2	$(1-e^{\lambda\Delta})/2$	8	2	1	2	1
	0	2	3	$(1-e^{\lambda\Delta})/2$	9	2	2	2	1
	1	0	1	$e^{-\lambda\Delta}$	10	3	0	3	1
4	1	0	2	$(1-e^{-\lambda\Delta})/2$	11	3	1	3	1
	1	0	3	$(1-e^{-\lambda\Delta})/2$	12	3	2	3	1

Table 2: The CPT of $P1\#$ and $P2\#$.

the four possible effects on H according to the current state of P1, P2 and V, as specified in Tab. 1. In particular, the value 0 represents the decrease of H, 1 represents the steadiness of H, 2 models the slow growth of H, 3 models the quick growth of H. There is no difference between the states OFF and S_OFF, or ON and S_ON, in terms of effects on H. The CPT of $trend$ reflects the content of Tab. 1. For instance, the first entry specifies that if all the components P1, P2 and V are OFF, then H is steady ($trend = 1$) with probability 1.

A DBN can represent discrete quantities in terms of the values of variables. H is a continuous measure to be discretized in order to be modelled by a DBN variable. On one hand, a low number of discrete intermediate levels may lead to some approximation of the inference results. On the other hand, a high number may increase in a relevant way the size of several CPTs and as a consequence, the complexity of the model analysis. In order to achieve a good trade-off between accuracy and complexity, in the DBN we discretize H into 13 intermediate levels. To this aim, we exploit the variable $Level$ whose value can vary between 0 and 12. This means that the distance between an intermediate level and the following one is 0.5 m: Tab. 3 defines the correspondence between the 13 values of $Level$ and the effective liquid level in the tank. Given that two consecutive intermediate values differ by 0.5 m, in the DBN we can represent the variation of H for the same quantity by increasing or decreasing $Level$ by one unit. If the variation rate for P1, P2 and V is Q=0.6 m/h (Section 3), then a variation of H by 0.5 m (1 unit for $Level$) due to the action of a single component, takes 0.8333 h of time. We set the time discretization step Δ to this value in such a way that $Level$ may change by 1 during one time step. The parameter Δ is used in the CPTs of $P1\#$ (Tab. 2), $P2\#$, $V\#$ to compute the component (un)reliability.

In the DBN, the variable $Level\#$ (current H) depends on $Level$ (H in the previous time step) and on $trend$ (the effect due to current state of P1, P2 and V). In particular, with respect to the value of $Level$, the value of $Level\#$ is the same if $trend = 1$, is decreased by 1 if $trend = 0$, is increased by 1 if $trend = 2$, or by 2 if $trend = 3$. All of this is specified in the CPT of $Level\#$ (Tab. 4).

$Level$	actual level	region	$Level$	actual level	region	$Level$	actual level	region
12	+3.0 m	2	7	+0.5 m	1	2	-2.0 m	0
11	+2.5 m	2	6	+0.0 m	1	1	-2.5 m	0
10	+2.0 m	2	5	-0.5 m	1	0	-3.0 m	0
9	+1.5 m	2	4	-1.0 m	0			
8	+1.0 m	2	3	-1.5 m	0			

Table 3: The values of $Level$ and the corresponding intermediate liquid levels (Figure 1.a).

$Level$	$trend$	***Level#***	prob.	$Level$	$trend$	***Level#***	prob.
0	0	0	1	7	0	6	1
0	1	0	1	7	1	7	1
0	2	0	1	7	2	8	1
0	3	0	1	7	3	9	1
1	0	0	1	8	0	7	1
1	1	1	1	8	1	8	1
1	2	2	1	8	2	9	1
1	3	3	1	8	3	10	1
2	0	1	1	9	0	8	1
2	1	2	1	9	1	9	1
2	2	3	1	9	2	10	1
2	3	4	1	9	3	11	1
3	0	2	1	10	0	9	1
3	1	3	1	10	1	10	1
3	2	4	1	10	2	11	1
3	3	5	1	10	3	12	1
4	0	3	1	11	0	10	1
4	1	4	1	11	1	11	1
4	2	5	1	11	2	12	1
4	3	6	1	11	3	12	1
5	0	4	1	12	0	12	1
5	1	5	1	12	1	12	1
5	2	6	1	12	2	12	1
5	3	7	1	12	3	12	1
6	0	5	1				
6	1	6	1				
6	2	7	1				
6	3	8	1				

Table 4: The CPT of $Level\#$.

The entries with $Level = 0$ and $Level = 12$ correspond to the dry out and the overflow respectively; the probabilities in such entries express the assumption that H does not change any more if a system failure condition is reached.

H can be inside one of three regions: H$\leq$HLA, HLA$<$H$<$HLB, H$\geq$HLB (Section 3). The variable $Level\#$ influences $region\#$ whose value can be 0, 1, or 2 in order to represent the above three regions respectively, and the corresponding commands (Section 3). The CPT of $region\#$ maps the values of $Level\#$ into the corresponding value of $region\#$ according to Tab. 3.

System states. We need the ternary variable $state\#$ to model the three possible states of the system: working, dry out or overflow. In particular, the working state is any situation where the dry out or the overflow has not occurred yet. These states are determined by H, so $state\#$ is influenced by $Level\#$. The value 0 of $state\#$ represents the dry out, the value 1 models the working state, and 2 indicates the overflow. In the CPT of $state\#$, we set this variable to 0 only when $Level\# = 0$ and we set it to 2 only when $Level\# = 12$, according to Tab. 3. In any other case, $state\#$ is set to 1. Since $Level\#$ does not change any more its value in case of dry out or overflow (as described above), $state\#$ maintains its value if set to 0 or 2.

6. Model analysis

Predictive results. First, we compute the *cumulative distribution function* for the dry out (cdf_{dry}) and the overflow (cdf_{ov}). The value of cdf_{dry} and cdf_{ov} at time t is the probability that the system has failed because of the dry out and overflow, respectively, during the time period $(0, t)$. In other words, cdf_{dry} and cdf_{ov} are the system unreliability because of the dry out and the overflow, respectively. These measures can be computed on the DBN models by means of the filtering task with an empty stream of observations (Section 4).

As in [1], the system is evaluated for a mission time varying between 0 and 1000 h. In DBN, the time is discrete (Section 4), and two consecutive time steps differ by the interval Δ which is set to 0.8333 h in the DBN models of the benchmark (Section 5). So, in order to evaluate the system from 0 to 1000 h, we have to inference the model from 0 to 1200 time steps. For example, the system evaluation at 400 h is given by the DBN analysis at 480 time steps (480 = 400 h / 0.8333 h).

At each time step, the variable $state\#$ is queried to obtain the probability distribution of its values 0, 1, 2, corresponding to the dry out, working, and overflow condition, respectively (Section 5). So, the probability that $state\#$ is equal to 0 at time t, provides cdf_{dry} at that time, while cdf_{ov} is given by the probability that $state\#$ is equal to 2.

The results returned by DBN analysis are quite similar to those obtained by the techniques described in Section 2, as shown in Tab. 5 (cdf_{dry}) and in Tab. 6

time	**step**	**DBN an.**	SAN sim.	GSPN an.	FSPN sim.
200 h	240	2.2789E-2	2.2390E-2	2.2077E-2	2.400E-2
400 h	480	6.6455E-2	6.5990E-2	6.5827E-2	6.730E-2
600 h	720	9.5366E-2	9.5290E-2	9.5014E-2	9.360E-2
800 h	960	1.1040E-1	1.1003E-1	1.1022E-2	1.084E-1
1000 h	1200	1.1777E-1	1.1747E-1	1.1768E-2	1.165E-1

Table 5: The cumulative distribution function for the dry out (cdf_{dry}).

time	**step**	**DBN an.**	SAN sim.	GSPN an.	FSPN sim.
200 h	240	1.9890E-1	1.9914E-1	1.9518E-1	2.0050E-1
400 h	480	3.6172E-1	3.6207E-1	3.5987E-1	3.6220E-1
600 h	720	4.3652E-1	4.3665E-1	4.3568E-1	4.4160E-1
800 h	960	4.6997E-1	4.7063E-1	4.6959E-1	4.7630E-1
1000 h	1200	4.8538E-1	4.8572E-1	4.8520E-1	4.9100E-1

Table 6: The cumulative distribution function for the overflow (cdf_{ov}).

(cdf_{ov}). This verifies that DBN analysis generates results with a good degree of accuracy. The differences in the results are due to the model evaluation approach (analysis or simulation), the modelling power of each formalism (DBN, SAN, GSPN, FSPN), and the assumptions holding in the model (Section 2). For instance, the DBN, the GSPN, and the SAN model capture variations of H by 0.5 m, 1 m and 0.01 m respectively, while H is a continuous variable in the FSPN. The DBN and the GSPN model undergo analysis, while the SAN and FSPN model are simulated.

Diagnostic results. DBN can be exploited to compute measures conditioned by observations (Section 4). In this case, the difference between a filtering and a smoothing inference (Section 4) relies on the fact that in the former case, while computing the probability at time t, only the evidence (observations) gathered up to time t is considered; on the contrary, in the case of smoothing the whole evidence stream is always considered in the posterior probability computation. For diagnosis purposes, filtering can be exploited to perform the on line diagnosis of the system. This means evaluating the state of components during the monitoring of the system behaviour. For instance, in our case study, if we assume that the value of H is observable at each time step t, then we can compute the probability of each possible state of P1, P2 and V at t. In this way, we can estimate the causes of the current value of H. Smoothing instead, may be exploited in order to reconstruct the history of the system components for a kind of temporal diagnosis. For instance, we may be interested in evaluating the probability of each state of P1, P2 and V at each time step, based on the observations about H, collected during all the system mission time. These kinds of measures were not computed in the previous works about the benchmark (Section 2). They are an additional value given by DBN.

In order to clarify these concepts, we provide an example of filtering and smooth-

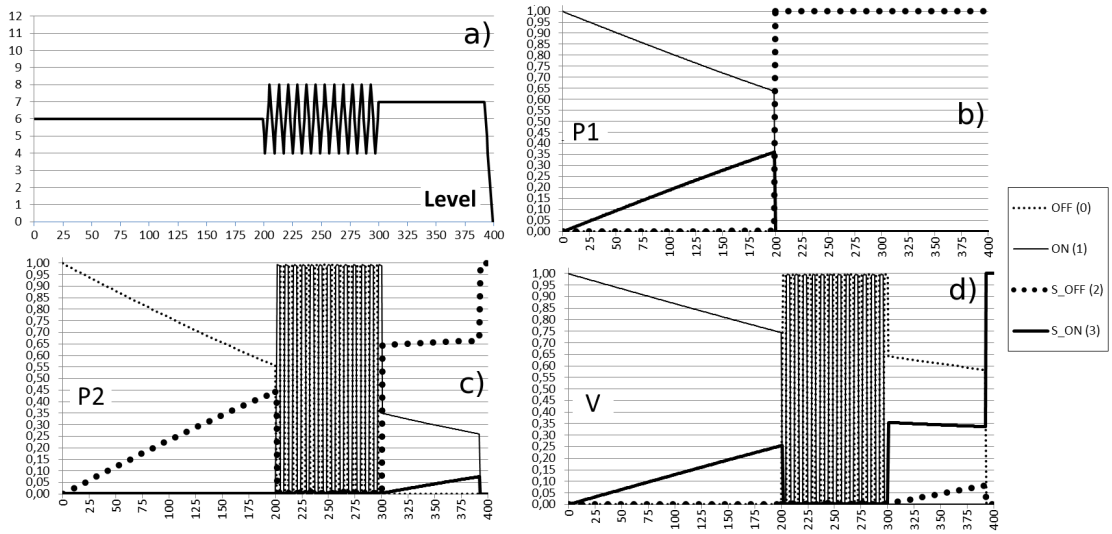


Figure 2: The observations about $Level$ and the diagnostic results given by the **filtering** task.

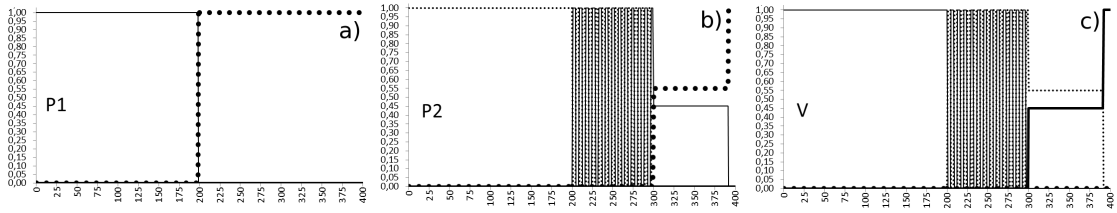


Figure 3: The diagnostic results given by the **smoothing** task.

ing applied to the DBN model, assuming to observe the value of the variable $Level$ at each time step; $Level$ represents H in the model (Tab. 3). The progress of $Level$ during the time is depicted in Figure 2.a. In order to detect if current H is due to a particular state of P1, P2 or V, we perform the filtering task on the DBN model, querying the variables $P1\#$, $P2\#$ and $V\#$, with the aim of computing the probabilities of their possible values (0, 1, 2, 3) corresponding to the possible states (OFF, ON, S_OFF, S_ON, respectively) of the components (Section 5).

The **filtering** results are depicted in Figure 2. $Level$ is observed steady to 6 from time step 0 to time step 199. So, initially, P1 is ON, P2 is OFF and V is ON with probability 1. This configuration is still possible from 1 to 199, but there is an increasing probability of S_ON for P1 and V, and S_OFF for P2, due to the observed steadiness of $Level$ and the failure rate of the components.

At time step 200, $Level$ becomes 5. The filtering results indicate that the cause of this decrease is certainly the S_OFF state of P1. At 201, $Level$ reaches 4 (H=HLA), so we expect an order from C with the aim of switching on the pumps and switching off V. Actually in this time step, P2 turns to ON, while V turns to OFF, with probability 1. $Level$ grows from 4 to 8 during the time steps from 201 to 205. In particular, at 205, due to $Level = 8$ (H=HLB), C

should order the pumps to switch off, and V to switch off. This is confirmed by the probabilities of P2 to be OFF and of V to be ON, both equal to 1 in this time step. This configuration of the components leads to a decrease of $Level$ from 8 to 4 during the time steps from 205 to 209. At 209, $Level$ is equal to 4 (H=HLA), so C successfully sets P2 to ON and V to OFF according to the filtering results. This configuration will determine another growth of $Level$ up to 8 (H=HLB) with another inversion of the states of P2 and V to decrease $Level$ again. This fluctuation of $Level$ lasts until the time step 300 and is reflected by the alternation of the states ON and OFF for P2 and V, during this time.

At 301, we observe that $Level$ is equal to 7, the same value observed at 300. In other words, $Level$ has interrupted its growing stage maintaining its value. The filtering results provide two alternative causes for this event, with different probabilities: P2 is S_OFF or V is S_ON. From 302 to 392 $Level$ maintains the value 7; because of this, during these steps, P2 may be ON or S_ON (combined with V in S_ON state), or S_OFF (combined with V in OFF or S_OFF); V instead, may be OFF or S_OFF (assuming that P2 is S_OFF), or S_ON (assuming that P2 is ON or S_ON). At 393, $Level$ becomes 6 (H is decreasing). The filtering task deduces that the certain cause is the contemporary S_OFF state of P1 and the S_ON state of V. This is confirmed in the next steps leading to the dry out.

The **smoothing** results for Scenario 2 (Figure 3) show a more precise diagnosis: the anticipated knowledge about the values of $Level$ excludes the possibility that P1, P2 or V may be S_OFF or S_ON between the time steps 1 and 199. Their state is certain during that period. The diagnosis between the time steps 200 and 300 is confirmed; the probability that P2 is S_ON and the probability that V is S_OFF between 301 and 392, are both null in the smoothing task results, while such states were possible according to the filtering output (Figure 2).

7. Conclusions

A benchmark on dynamic reliability taken from the literature has been evaluated. The predictive results about the system unreliability that we obtained, are in general quite similar to those computed by means of other techniques. This proposes DBN as suitable models to deal with dynamic reliability cases, with two main advantages: **1)** with respect to state space based models, the use of a DBN takes advantage of factorization of the system state space into the model variables. **2)** DBN introduce the possibility of computing measures conditioned by observations, at specific times. This has been applied in order to compute diagnostic measures about the state of components.

References

[1] Marseguerra, M. and Zio, E., Monte Carlo Approach to PSA for dynamic process system, *Reliability Engineering and System Safety*, **52** 227–241 (1996).

[2] Sahner, R., Trivedi, K. and Puliafito, A., *Performance and Reliability Analysis of Computer Systems*, Kluwer Academic Publisher (1996)

[3] Langseth, H. and Portinale, L., Bayesian Networks in reliability, *Reliability Engineering and System Safety*, **92** [1] 92–108 (2007).

[4] Murphy, K., *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, UC Berkley (2002).

[5] Weber, P. and Jouffe, L., Reliability modelling with dynamic Bayesian networks, in *Proceedings of the Symposium on Fault Detection, Supervision and Safety of Technical Processes*, Washington DC, USA (2003).

[6] Portinale, L., Codetta-Raiteri, D. and Montani, S., Supporting Reliability Engineers in Exploiting the Power of Dynamic Bayesian Networks, *International Journal of Approximate Reasoning*, **51** [2] 179–195 (2010).

[7] Codetta-Raiteri, D. and Bobbio, A., Solving Dynamic Reliability Problems by means of Ordinary and Fluid Stochastic Petri Nets, in *Proceedings of the European Safety and Reliability Conference*, pp 381–389, Gdansk, Poland (2005).

[8] Codetta-Raiteri, D., Modeling and simulating a benchmark on dynamic reliability, as a Stochastic Activity Network, in *Proceedings of the European Modeling & Simulation Symposium*, pp 545–554, Rome, Italy (2011).

[9] Gribaudo, M., Sereno, M., Horvath, A. and Bobbio, A., Fluid Stochastic Petri Nets augmented with flush-out arcs: Modelling and analysis, *Discrete Event Dynamic Systems*, **11** [1/2] 97–117 (2001).

[10] Sanders, W. and Meyer, J., Stochastic activity networks: Formal definitions and concepts, *Lecture Notes in Computer Science*, **2090** 315–343 (2001).

[11] Codetta-Raiteri, D. and Portinale, L., Bayesian networks applied to dynamic reliability with predictive and diagnostic purposes, *submitted to Reliability Engineering and System Safety* (2012).

[12] Portinale, L., Bobbio, A., Codetta-Raiteri, D. and Montani, S., Compiling dynamic fault trees into dynamic Bayesian nets for reliability analysis: The Radyban tool, *CEUR Workshop Proceedings*, **268** (2007).

Condition monitoring data in the study of offshore wind turbines' risk of failure

Maria C. Segovia[1], Matthew Revie[1], Francis Quail[2]

[1]Department of Management Science, University of Strathclyde, Glasgow
[2] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow

Abstract

Unplanned maintenance actions entail a period of inactivity of wind turbines and therefore a loss of revenues. This is even more pronounced in the case of offshore wind farms because of difficulties in access. To this end, condition monitoring (or health monitoring) systems have been implemented on wind turbines by manufacturer to support maintenance decision making by operators. However, a major concern with using condition monitoring is the creation of false positives.

In this paper, we have considered how SCADA data may be used to support reliability and maintainability models. We have taken into account SCADA data to assess the level of degradation of a gearbox at any instant of time. Workshops with offshore engineers suggested that degradation is dependent on a small number of variables, e.g. turbulence intensity, wind speed, temperature, etc., and as such, these information sources have been considered in our model to support maintenance decision making.

Dynamic Bayesian Belief Networks allows for modelling multiple information sources (including expert judgement) and dynamic phenomena, e.g. system deterioration. They are also capable of representing dependencies between variables of interest. For these reasons, we developed a Dynamic Bayesian Belief Network to assess the risk of failure of the component under study at any instant of time. We considered the information provided by SCADA about the factors affecting the degradation of the system. In this way, we obtained an estimation of the risk of failure of the system.

1. Motivation

The wind power industry has grown considerably over the past 15 years. By the end of 2011 it was able to produce 204 TWh of electricity and meet 6.3% of the EU's total electricity demand (Arapogianni et al., 2012). The European Wind Energy Association (EWEA) expects 230 GW of installed capacity in 2020 and 400 GW by 2030. As a result, wind power is establishing itself as the main power technology in the EU (Arapogianni et al., 2012). Additionally, in 2010 2,965 MW of wind capacity was operational offshore, equating to 3.5% of the total installed wind energy capacity. This share is expected to increase to 17.4% by 2020 (40 GW) and 37.5% by 2030 (150 GW) (Arapogianni et al., 2012).

Offshore wind farms pose different challenges to onshore wind farms with respect to reliability and maintenance. Due to operating in harsher climates, the need to use specialised vessels, large losses associated with down time and greater technical uncertainty from using wind turbines without a significant operating history, offshore operation and maintenance costs are greater; in some cases as much as 25%-30% of the cost of the energy, (Nielsen and Sørensen (2010)). For this reason, it is of paramount importance to prevent failure and reduce the unavailability of the wind turbines. To this end, condition monitoring (or health monitoring) systems are currently being used by the wind industry.

The core of a condition monitoring system (CMS) are fault diagnosis algorithms that provide early warnings about the occurrence of mechanical and electrical faults. The aim of CMSs is to predict the failure of major components and from this, repairs actions can be planned prior to failure. This is particularly important in offshore wind farms, where weather constraints can delay a repair action for several weeks. The information provided by CMSs is starting to be considered to support reliability, maintainability and logistic models, however, as yet, it not clear what the economic benefit of these systems are (McMillan and Ault, 2008).

The aim of this paper is to explore how a particular methodology, Dynamic Bayesian Networks, can be used to model reliability and maintenance decisions taken during the lifetime of a wind farm. The typical decisions we wish to support are what kind of maintenance action we should perform, i.e. inspection, repair, replace, and when it should be done, i.e. best scheduled of the intervention.

Bayesian Belief Networks (BBNs) are frequently used to model systems subject to uncertainty. BBNs are probabilistic graphical models that are able to represent the dependencies between the variables of interest. BBNs have been applied extensively in system reliability and maintenance modelling, (Pearl (1988), Neil et al. (2001), Sigurdsson et al. (2001), Langseth and Lindqvist (2003), Weidl et al. (2003) and Langseth and Portinale (2006). For a detailed description of BBNs, both theoretical and applied, see, for example, Pearl (1988), Lauritzen (1996), Cowell et al. (1999), Jensen (2001).

One drawback of BBNs is that they are static models that represent the joint probability distribution at a fixed point or interval of time. As such, they are unable to capture the dynamic behaviour of a system degrading through time due to usage. When modelling systems' reliability, whose nature is dynamic, Dynamic Bayesian Belief Networks (DBBNs) are more suitable. DBBNs are an extension of BBNs that allow for the modelling of dynamic phenomena, i.e. the system deterioration and the weather, and allow us to consider temporal dependencies. DBBNs are, in this way, a useful tool to support temporal-decision making. Recently, McNaught and Zagorecki (2009) applied DBBNs for prognostic modelling of equipment in order to better inform maintenance decision-making. They consider CMS to infer the true condition of the system and also external factors that accelerate the wear-out of the system.

In this paper, we consider how DBBNs can be used to model the external factors affecting the deterioration of the system by taking into account the

information provided by SCADA systems. We anticipate using this model to support maintenance decision making and to measure the impact of these decisions on the reliability of the system. We focus in assessing the risk of failure of one of the major components of the wind turbine, i.e. the gearbox. The gearbox is a critical component whose failure implies the longest downtime compared to other components, (Ribrant and Bertling, 2007).

In spite of the similarities to the paper of McNaught and Zagorecki (2009) there are some important differences. While they consider just discrete variables, we have developed a hybrid DBBN, with continuous and discrete variables. In addition, for the purposes of this research, we restrict ourselves to modelling using a Kalman Filter (KF). The KF can be seen as a special kind of DBBN where the joint probability distribution is Normal. The aim of our model is to estimate the deterioration state of the system.

This paper is organized as follows: Section 2 provides an overview of DBBNs and the KF. In Section 3 we describe the structure of the DBBN adopted in this problem, i.e. the selected factors specified by experts. The population of the model is discussed in Section 4. In Section 5 the output of the analysis is illustrated. Conclusions and future work are provided in Section 6.

2. Methodology

2.1 *Dynamic Bayesian Networks*

BBNs were introduced in the 1980s, (Pearl (1988)), as a flexible and powerful probabilistic modelling framework that makes them suitable for applications in the field of reliability and maintenance. They are a tool for reasoning under uncertainty and allow dependencies between variables of interest to be modelled. Furthermore, as the statistical methodology is Bayesian, data can be combined with expert opinion, e.g. considering how different environmental factors or design considerations will affect reliability.

A BBN is a compact representation of a multivariate statistical distribution. It encodes the probability density function governing a set of n random variables $X = (X_1, X_2, \ldots, X_n)$ by specifying a set of conditional independent statements together with a set of conditional probability functions, (Langseth and Portinale (2006)). More specifically, a BBN can be seen as a static graphical model, in which the nodes represent random variables (continuous or discrete), and the edges imply direct dependencies between the linked variables (see Figure 1). The strength of these dependencies is quantified by conditional probabilities.

The joint probability distribution of the variables is determined taking into account the BBN structure (Figure 1), using the chain rule, and considering the conditional independence assumptions. The joint distribution of a set of variables $X_1, X_2, \ldots, X_n$ is given by the following expression,

$$f\left(x_1, x_2, \ldots, x_n\right) = \prod_{i=1}^{n} f\left(x_i \middle| pa\left(X_i\right)\right) \tag{1}$$

where $pa(X_i)$ denotes the set of all the parents of node X_i.

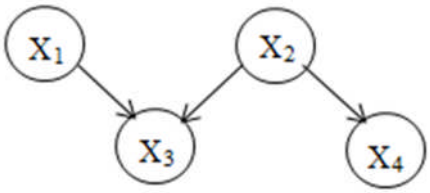


Figure 1: A simple Bayesian Belief Network

Nevertheless, BBNs are static models that represent the joint probability distribution at a fixed point or interval of time. To account for temporal dependencies, we need an explicit representation of time in a BBN. DBBNs extend BBNs to allow for reasoning in a dynamic world where changes occur over the time. In a DBBN, we consider successive time instants and a random variable, i.e. a node, is associated with each time instant. A DBBN captures the evolution of a system over time by interconnecting static BBNs over time slices (Wang et al., 2010). For more information, see Dean and Kanazawa (1989) and Murphy (2002).

Assuming that, $X_t = (X_1(t), X_2(t), \ldots, X_n(t))$ represents the set of random variables at time t, we have that the joint distribution in a DBBN is given by the following expression,

$$f\left(x_{1:T}\right) = \prod_{t=1}^{T} f\left(x_1(t), \ldots, x_n(t)\right) = \prod_{t=1}^{T}\prod_{i=1}^{n} f\left(x_i(t) \middle| pa\left(X_i(t)\right)\right) \tag{2}$$

where $X_{1:T}$ denotes the corresponding variables in each time slice, for some $T > 0$, see Zitrou et al. (2010).

2.2 *Kalman Filter*

The KF can be seen as a DBBN where the Conditional probability distributions and the joint distribution are Gaussian (Murphy (2002)). The KF is an optimal recursive data processing algorithm that uses a series of measurements observed over the time and produces estimations of unknown variables. The KF uses (1) the knowledge of the system and measurements device dynamics (2) the uncertainty in the dynamics models and (3) any initial conditions of the variables of interest (Maybeck (1979)). The algorithm works in a 2 step process. First, the KF produces estimations of the current state of the variables of interest along with their uncertainties. Second, once a measurement is observed, the previous estimations are updated using a weighted average, with more weight being given to estimations with higher certainty.

Formally, we can define the KF in the following way (Welch and Bishop(2001)):

The Kalman filter addresses the general problem of trying to estimate the state $X_t \in \Re^n$ of a discrete-time controlled process that is governed by the linear stochastic difference equation

$$X_t = AX_{t-1} + BU_t + V_t \tag{3}$$

with a measurement $Z_t \in \Re^m$. That is

$$Z_t = HX_t + W_t$$

The random variables V_t and W_t represent the process and measurement noise (respectively). The matrix A relates the state at the previous time step $(t-1)$ to the state at the current step t, in the absence of either a driving function or process noise. The matrix B relates the optional control input U_t to the state X_t. The matrix H in the measurement equation relates the state X_t to the measurement Z_t.

KF has been broadly applied. In particular, Qu and Hahn (2009) compare different kinds of KF for detection of abnormal operating conditions in industry.

3. Model approach

Mechanical systems are subject to aging, which entails sooner or later to the deterioration of their performances and ultimately their failure. External conditions can accelerate this degradation and cause unexpected failures. For an offshore wind farm, an unplanned maintenance action can suppose a considerable loss of revenues. CMS can help us to prevent the failure; however, a major concern when using CMS is the creation of false positives.

3.1 *DBBN structure*

As a first approach to understand how the gearbox of an offshore wind turbine deteriorates, several interviews with engineers were conducted. During these interviews factors influencing the deterioration of the system were discussed. From this, the variables that were considered to have the biggest impact on the reliability of the gearbox were *Turbulence intensity,* defined as the ratio of the Wind Speed Standard deviation and the mean wind speed determined from the same set of measured data samples of wind speed, and taken over a specified period of time (IEC 61400-1 (1999))*, Generator rpm* and the *Maintenance decisions*.

In our approach, the deterioration of the gearbox is measured in terms of a variable called *Cumulated Effective Number of Rotations of the Generator (CENRG).* We use *CENRG* as we believe that the measurement of the deterioration through the usage of the gearbox is a more accurate representation of the deterioration than the calendar age. The variable *CENRG* depends on the *Generator rpm,* which is an indicator of the usage of the generator, and on the *Turbulence intensity*, which provides the conditions of usage. The latter variable corresponds to the external conditions accelerating the deterioration on the gearbox. For wind turbines exposed to high turbulence intensity, it is expected that the gearbox will deteriorate much faster than when this intensity is low.

We also consider the impact of *Maintenance decision*s. While the first two variables, *Generator rpm* and *Turbulence intensity,* increase the deterioration on the gearbox, the maintenance action has an opposite impact on the deterioration.

We expand *Turbulence Intensity* and *Generator rpm* to include *Observed Turbulence Intensity* and *Observed Generator rpm*. The reason is that the SCADA system is providing evidence of the *True Turbulence Intensity* and the *True Generator rpm* every 10 minutes. Applying KF, we are able to infer the *True Turbulence Intensity* and the *True Generator rpm*. Therefore, in our DBBN, the time slices correspond to 10 minutes intervals. The DBBN can also predict the evolution of the system considering the previous history of the gearbox.

The dynamic evolution of the system is represented by means of a DBBN given by Figure 2. The arcs on the figure correspond to direct probabilistic dependencies between the different variables. Straight arrows indicate relationships within the same time slice. Circular arrows indicate a dependence across time slices, in this case from one time slice to the next one.

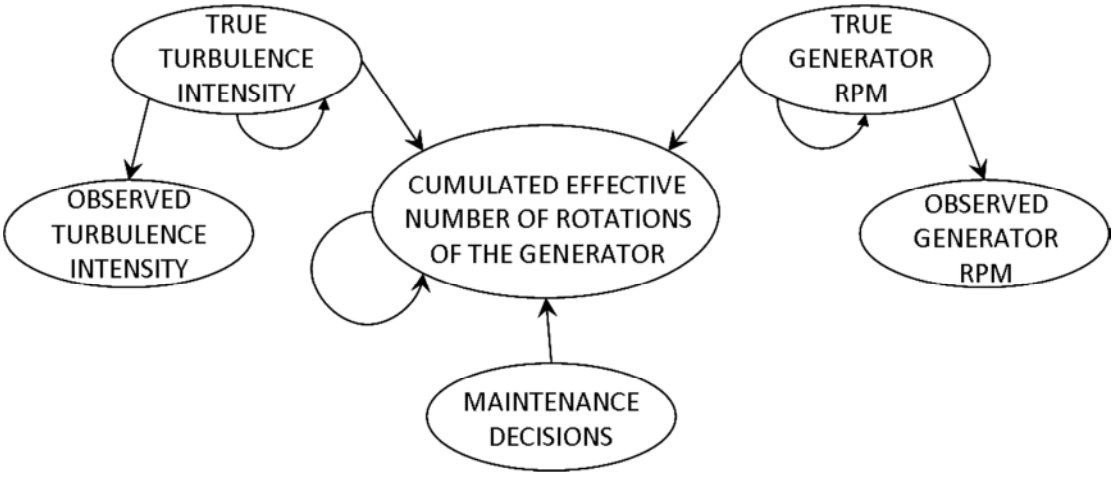


Figure 2. DBBN for the deterioration of the gearbox

3.2 *Obtaining the gearbox deterioration*

Considering the structure of the DBBN, we are able to infer the deterioration level of the gearbox in the following way. First, we consider how the *True Turbulence Intensity* is inferred by the *Observed Turbulence Intensity* through the KF,

- *KF for Turbulence Intensity*

$$TI_{t+1} = A_{TI} TI_t + VTI_{t+1} \tag{4}$$

$$O_{TI_{t+1}} = H_{TI} TI_t + WTI_{t+1}$$

$O_{TI_{t+1}}$ is the observation at time slice $t+1$ of the *True Turbulence Intensity*, TI_{t+1}, this is obtained through the SCADA system, being A_{TI}, H_{TI} matrices. VTI_{t+1} and WTI_{t+1} represent the process and measurement noise respectively (see Section 2.2).

We do the same for *True Generator rpm* and the *Observed Generator rpm*

- KF for Generator rpm

$$Grpm_{t+1} = A_G Gprm_t + VG_{t+1} \quad (5)$$

$$O_{Grpm_{t+1}} = H_G Grpm_t + WG_{t+1}$$

$O_{Grpm_{t+1}}$ is the observation of the *True Generator rpm*, $Grpm_{t+1}$, provided by SCADA at time slice $t+1$. A_G, H_G are matrices. VG_{t+1} and WG_{t+1} represent the process and measurement noise respectively (see Section 2.2).

Taking this into account we can estimate the *CENRG* as follows*:*

$$CErpm_{t+1} = (CErpm_t + \gamma_{t+1} Grpm_{t+1})(1-(\rho_{t+1}) \quad (6)$$

Where,

- $CErpm_{t+1}$ and $CErpm_t$ represent the *CENRG* at time slices $t+1$ and t respectively.
- $Grpm_{t+1}$ is the *Generator rpm* between the time slices t and $t+1$.
- $\gamma_{t+1} \geq 1$ is the *True Turbulence Intens*ity between time slices t and $t+1$. This parameter corresponds to the impact of the turbulence on the deterioration of the gearbox. If the *True Turbulence intensity* in that interval exceeds a threshold, γ_{t+1} is greater than 1, so the number of *Generator rpm* in that interval is multiplied, accelerating the deterioration of the gearbox. The impact value and the different thresholds are elicited from engineers.
- $0 \leq \rho_{t+1} \leq 1$ represents the effectiveness of the maintenance action performed, if any, at time slice $t+1$. When no maintenance action takes place, this value is 0. If there is an intervention, then $\rho_{t+1} > 0$. This value depends on the intervention carried out, e.g. if we replace the gearbox ρ_{t+1} is 1, i.e. we reset the CENRG to 0. The different values of ρ_{t+1} depend on the maintenance action. As we can see, maintenance actions remove partially or completely the cumulated damage on the system during its whole life. Our approach is inspired on the Effective age models known as *Arithmetic Reduction of the Age (ARA models).* For a detailed explanation of these models see Doyen and Gaudoin (2004).

Once the *CENRG* is calculated, the probability of failure is estimated using a Weibull distribution, this was decided after discussion with engineers. The probability of failure is given in terms of the *CENRG*. The probability of failure at time slice *t* becomes:

$$F\left(CErpm_t\right)=1-e^{-\left(\frac{CErpm_t}{\lambda}\right)^k},\ \ CErpm_t \geq 0 \qquad (7)$$

where $k>0$ is the shape parameter and $\lambda>0$ is the scale parameter. We have estimated the parameter using experts elicitation

4. Validation and population of the DBBN

4.1 Gathering some data

As explained in previous sections, we have historical SCADA observations. In particular we have observations of *Wind Speed average, Standard deviation of the Wind Speed.* From these observations, we obtain the *Observed Turbulence Intensity.* We have also observations of *Generator rpm.* In order to populate the model, we must specify the joint probability distribution across all the variables. These values have been elicited from engineers.

4.2 Quantitative Elicitation

We define expert judgement as "any structured method of acquiring knowledge from experts" (Bedford and Cooke, 2006). There are numerous examples of risk and reliability projects where expert judgement has been adopted.

We use expert judgement to structure and populate the model in this project due to the lack of available data. The model of the deterioration of the gearbox has been developed after discussion with engineers. During these interviews other variables were suggested, e.g. Wind Speed, Humidity, Dust, Load, Power output, Wind direction, Downtime, etc. Some of these variables will be included in a future version of the model.

Due to the lack of data, the distribution of the lifetime of the gearbox in terms of the CENRG was elicited from experts and fit to a Weibull distribution. The *Turbulence intensity impact* on the deterioration, the thresholds related and the *Maintenance actions impact*s were also elicited. We have considered three categories for the turbulence intensity, '*Normal*', '*Medium*' and '*High*', and we have associated different impacts to each category. For the Maintenance actions, we have distinguished between '*Inspection/Small repair*', '*Repair*', '*Replacement of the gearbox*', '*De-rating*' and '*No action*'. We have historical data of failures and repairs available so they have been used to partially validate the values obtained through a questionnaire.

5. Outputs and decision support

Once the model was populated, the data was used to forecast future deterioration beyond time step t. From the proposed DBBN we obtain a probability of failure at any instant of time. To do this, we use the historical dataset of observations and apply the KF to update our beliefs, i.e. to estimate the true value of these two variables. From this, we can then use the DBBN to infer the *CENRG*. This allows us to estimate the current probability of failure of the gearbox. We can also use the KF to predict the *Turbulence Intensity*

and the *Generator rpm* over the next time steps. By doing so, we can predict the future probability of failure over those time steps.

If we do not take into account CMS' evidence, the DBBN offers interesting outputs, i.e. the level of deterioration at any instant of time considering external factors and maintenance actions impact. In this way a maintenance action can be advised even before the CMS indicates a problem in the system. The understanding of the deterioration mechanisms of the gearbox can be used to review the company's policy concerning scheduled preventive maintenance actions and replacements, and in this way, to reduce costs.

5.1 *Illustration of outputs*

Due to the confidentiality of the data used during the development of the model, we use fictional data to illustrate the performance of the DBBN. Let us suppose that the distribution of the lifetime of the gearbox in terms of the *CERNG* is gven by a Weibull distribution with the shape parameter $k = 1.5$ and the scale parameter $\lambda = 1800000000$. Consider prior distributions over the *True Turbulence Intensity* and *Generator rpm* as Normal distributions with parameters $\mu_{TI} = 0.2, \sigma_{TI} = 0.5$ and $\mu_G = 1000$, $\sigma_G = 100$ respectively. In addition, we elicit $A_{TI} = 1.001, H_{TI} = 1, A_G = 1,$ and $H_G = 1.001$. We also provide the noises distributions for each KF, however, due to space restrictions, they are not provided here.

Suppose that we have SCADA data of *Turbulence intensity* and *Generator rpm* since the wind turbine started its operation as new. Taking into account the previous distributions and the structure of our DBBN, we are able to infer the survival probability at the present moment and also in the near future, as we can see in Figure 3.

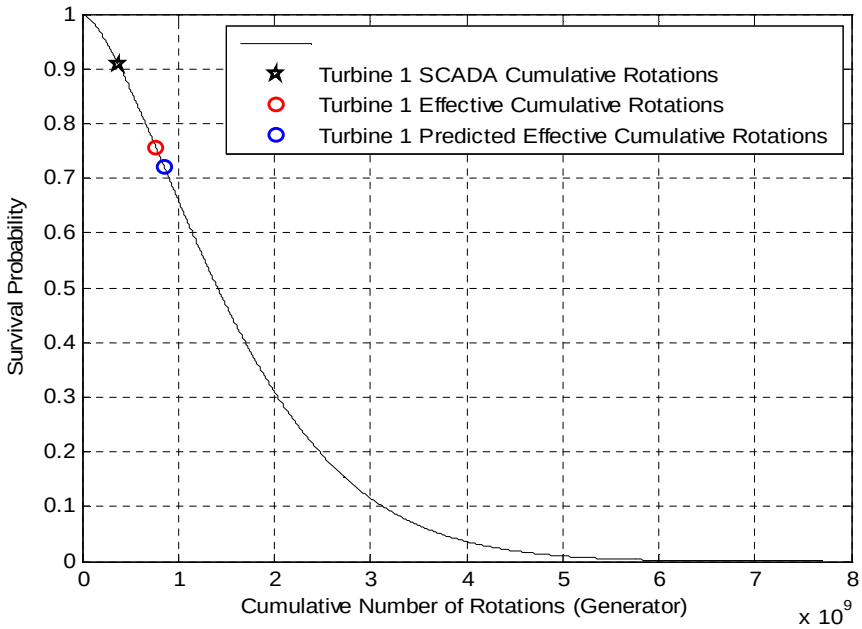


Figure 3. Survival probabilities for Turbine 1

In Figure 3, the black star represents the survival probability if we do not consider the impact of the *Turbulence Intensity* over the system. The red

circle is the Survival probability if we consider the impact of the external factors, i.e. the Turbulence Intensity. The blue circle is the prediction of the survival probability in 90 days. This modelling allows operators of wind turbines to compare the 'health' of multiple wind turbines.

We also can estimate the impact on the Survival probability function after the different maintenance actions that can be performed on the gearbox. We assume that the repair action recovers the system 50%, i.e. the *CENRG* decreases 50%. Replacement recovers the system 100%, The rest of the maintenance actions considered, i.e. 'Inspection/Small repair', 'De-rating' and 'None', keep the same level of degradation.

In Figure 4 we consider a degrade gearbox and estimate the effect of the different maintenance actions over the Survival probability of the system.

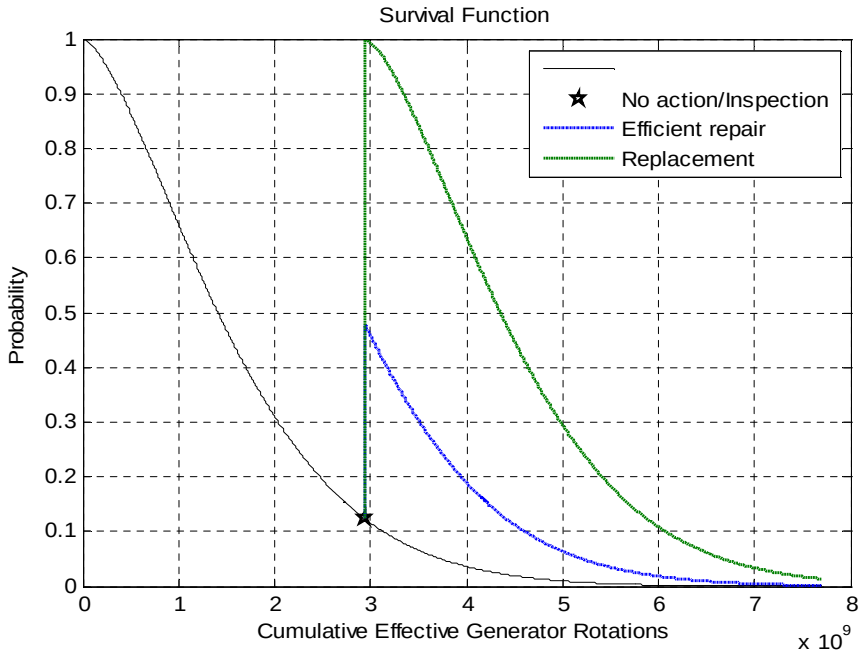


Figure 4: Maintenance actions' effects on the Gearbox's Survival probability

6. Conclusions and future work

In the present paper we propose a DBBN for the estimation of the degradation of the gearbox of an offshore wind turbine. As we mentioned, the model can bring some benefits by its own, such as reduction of the maintenance costs and downtime with a better understanding of the deterioration process. This model constitutes a first approach and in the future, additional variables will be included. We will also consider the main components on the gearbox, the dependencies between those components and the dependencies between the different wind turbines within the wind farm.

In addition, we wish to include information provided by CMSs. Currently, machine learning algorithms are being applied to historical data to develop diagnostic and prognostic models. In future, this information will be incorporated within the DBBN to estimate the deterioration level of the gearbox. Besides, using the CMS data, we can estimate the impact of the

maintenance action over the system, i.e. if the maintenance action is not performed as expected, the failure rate will increase quicker after the intervention. CMSs can help us to reduce the uncertainty on the maintenance action impact as new data are available.

References

1. Arapogianni, A., Moccia, J. and Wilkes, J. *The impact of wind energy on jobs and the economy*. EWEA, Brussels, (2012).

2. Nielsen, J.J. and Sørensen J.D. *Bayesian networks as a decision tool for O&M of offshore wind turbines*. In Proceedings of the 5th international ASRANet Conference, (2010).

3. D. McMillan and G. Ault, "Condition monitoring benefit for onshore wind turbines: sensitivity to operational parameters," *Renewable Power Generation, IET,* vol. 2, pp. 60-72, (2008).

4. Pearl, J. *Probabilistic reasoning in intelligent system: Networks of plausible inference*. Morgan Kaufmann, CA, (1988).

5. Neil, M., Fenton, N., Forey, S. and Harris, R. Using Bayesian Belief Networks to Predict the Reliability of Military Vehicles. *IEE Computing and Control Engineering*, Vol 12 [1], pp 11-20 (2001).

6. Sigurdsson, J.H., Walls, L.A. and Quigley, J. Bayesian Belief Nets for managing expert judgment and modelling reliability. Quality and reliability engineering international. Vol 17, pp 181-190, (2001).

7. Langseth, H. and Lindqvist B.H. *A maintenance model for components exposed to several failure modes and imperfect repair*. In Mathematical and statistical methods in reliability, quality, reliability and engineering statistics. Doksum, K., Lindqvis, B.H., Singapore: World Scientific, pp 415–430, (2003).

8. Weidl, G., Madsen, A.L. and Dahlquist, E. *Object oriented Bayesian network for industrial process operation*. In: Proceedings of the first Bayesian applications modelling workshop. Agosta, J.M., Kipersztok, O., Laskey, K.B., Przytula, K.W. and Rish, I. editors, (2003).

9. Langseth, H. and Portinale, L. Bayesian networks in reliability. Reliability engineering and system safety. Vol 92, pp 92-108, (2006).

10. Lauritzen, S. L. *Graphical Models*. Oxford: Clarendon, (1996).

11. Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York, (1999).

12. Jensen, F.V. *Bayesian networks and decision graphs.* Springer-Verlag, New York, (2001).

13. McNaught, K.R. and Zagorecki, A. *Using dynamic Bayesian networks for prognostic modelling to inform maintenance decision making*. Proceedings of the IEEE Conference on Industrial Engineering and Engineering Management, Hong Kong, IEEE, Piscataway, NJ, pp. 1155-9, (2009).

14. Ribrant, J. and Bertling, L. M. Survey of Failures in Wind Power Systems With Focus on Swedish Wind Power Plants During 1997–2005. *IEEE Transactions On Energy Conversion*, Vol 22, pp 167-173, (2007).

15. Wang, R., Ma, L., Yan, C. and Mathew, J. *Preliminary study on bridge health prediction using Dynamic Objective Oriented Bayesian Network (DOOBN)*. In Proceedings of WCEAM 2010, Fifth World Congress on Engineering Asset Management, World Congress on Engineering Asset Management, (2010).

16. Dean, T. and Kanazawa, K. A model for reasoning about persistence and causation. *Computational Intelligence*, Vol 5, pp 142-150, (1989).

17. Murphy, K. *Dynamic Bayesian Network: Representation, Inference and Learning*. PhD thesis, U.C. Berkeley, (2002).

18. Zitrou, A., Bedford, T. and Walls, L. *A model for supporting decisions regarding the operation and maintenance of offshore wind turbines*. In Proceedings of the European Safety and Reliability Conference 2010 (ESREL 2010), (2010).

19. Maybeck, P.S. *Stochastic models, estimation, and control*. Academic Press, New York, (1979).

20. Welch, G and Bishop, G. *An introduction to the Kalman Filter*. Technical Report TR 95-041, University of North Carolina at Chapel Hill Department of Computer Science, (2001).

21. Qu, C. and Hahn, J. Process monitoring and parameter estimation via unscented Kalman filtering. *Journal of Loss Prevention in the Process Industries*, Vol 22, pp 703–709, (2009).

22. IEC 61400-1: “Wind turbine generator systems – Part 1: Safety requirements”, (1999);

23. Doyen, L. and Gaudoin, O. Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliability Engineering and System Safety*, Vo 84, pp 45-56, (2004).

24. Bedford, T. and Cooke, R. *Probabilistic Risk Analysis: Foundations and Methods*: Cambridge University Press, (2001).

Risk and Reliability: An Evolutionary Biologist's Perspective

Sara L. Goodacre
School of Biology, University of Nottingham, Nottingham, UK

Abstract

Evolutionary biologists study the process through which organisms adapt and survive. At the core of this process lies the generation of variation upon which natural selection acts. This can be viewed as an exploration of the different solutions that are possible for the same challenge (*i. e.* survival in a particular environment). The range of solutions that is explored is rarely if ever exhaustive, being constrained by the time that an organism has had to adapt, and by the starting point, which is itself a product of previous evolutionary processes.

There are parallels between the evolutionary process described above and the search for optimum solutions in engineering designs. Survival (ie. non-failure of an engineered design) is maximised by searching for the optimal solution given known parameters. The search may or may not have been exhaustive and the 'solution' adopted is the best set of conditions found from those searched. There is a difference, however, in that evolution can and regularly does explore options of high risk whereas engineered solutions may not.

A study has been made of the literature on the evolution of bacterial and invertebrate genomes to ask to what extent the most successful solutions found by evolution to the challenge of survival favour redundancy, diversity or repair, or a combination of each of these.

Long-term asset maintenance optimization at Scottish Water

Travis Poole[1], Tom Archibald[2], Jake Ansell[2] and Robert Murray[1]
[1]Scottish Water plc, Castle House, Dunfermline, Scotland, UK
[2]The University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh, Scotland, UK

Abstract

This paper reports on a Knowledge Transfer Partnership between Scottish Water and the University of Edinburgh which aims to produce an asset maintenance optimization tool which produces long-term (25-year) capital maintenance plans for Scottish Water's physical assets. Theoretical models are presented alongside the practical issues of producing maintenance plans for a series of assets. The suggested model for implementation is a Whittle Index model with a common currency for risk and costs. An example of its use is given along with an indication of sensitivity.

1. Introduction

Scottish Water is responsible for supply of clean water and removal of waste water for all of Scotland. These two activities are sometimes described as 'source to tap' and 'sink to sea'. This is achieved through 97,000 km of pipes, 1837 Waste Water Treatment and 297 Water Treatment Works. Scottish Water has a turnover approximately £1 billion. It wishes to maintain its perception amongst the Scottish people as "Scotland's most valued and trusted business. One we can all be proud of." To do this it needs to ensure that it delivers a good service at a fair price to its customers, the people of Scotland. Hence it seeks to have the 'best' feasible maintenance strategy.

Achieving the goal of a well-planned and efficient service in such a large organisation is not a simple task. Scottish Water has in the past developed a strong reactive culture which means that it is great at "fighting fires" (i.e., reacting quickly and efficiently to correct assets failures when they occur), but appreciates that it is necessary to move to the next stage which is being proactive in its maintenance strategy (i.e., reducing the number of asset failures by performing maintenance before the failures occur). The problem with in the reactive strategy is that it proves to be an inefficient use of capital funds, as reactive work tends to be more expensive than proactive. Scottish Water has recognised that it needs better planning both in the short-term (1-5 year) and long-term (~25 years). It needs to better understand the uncertainty of future requirements for maintenance activity and hence the maintenance budget required.

In addressing this requirement Scottish Water entered into a Knowledge Transfer Partnership (KTP) with the University of Edinburgh and University of Strathclyde. University of Edinburgh would be responsible for developing with Scottish Water models for 'optimal' maintenance strategies whilst University of Strathclyde would produce better deterioration models for Scottish Waters'

equipment. This paper will focus on the development of models for 'optimal' maintenance strategy primarily for planning purposes. Optimal maintenance strategy in this context means production of a feasible enhanced maintenance strategy. Whilst theoretically one can produce optimal solutions these are dependent on information and resources available and as usual it is assumed that there are limited resources and limited available data on which to build the model.

There are a number of models for deciding the optimal strategy for a single asset. Ansell et al (2004) and Archibald et al (2004) provided optimal models for individual assets with multiple maintenance actions. For multiple assets, though, it becomes more difficult given size of the problem that arises. Added to this is the fact that often the resources are limited and this constrains the problem. A number of approaches have been developed for such context and may be deployed in combination to achieve a 'best' solution which is close to optimal at reasonable computational effort. For example one might use the optimal model for each asset and then use the 'importance' of each asset to decide the priority for action. It could be cost based, age based, operation based or condition based. This strategy may mean that the maintenance action for some assets may be delayed, see Ansell et al (2011) for the impact of such delays. The proposal within this paper is the use of the Normalised Whittle Index Heuristic. As this approach may not always be applicable so the paper will also describe some alternative methods.

The structure of the paper is as follows. The next section will discuss the background of the study within the water industry. The subsequent section will detail the development of the model from Markov Decision Process to Whittle Index Heuristic, then the Normalised Whittle Index Heuristic, and includes the detail of alternative models. An application is presented as an illustration. This is followed by the conclusion.

2. The Context

One issue for all UK Water Companies is the length of the development time of the industry and that it was done on a local basis. Heather and Bridgeman (2007) encapsulate this in the following quote 'The water industry has developed over a long time into a large complicated network of assets'. As indicated in the introduction Scottish Water is a large company with many assets and each main asset will often be unique in some way. Certainly the equipment used at an asset may vary dependent on the date of installation. Even similar pieces of equipment will have experienced different maintenance regimes and environmental conditions. The complexity and scale facing Scottish Water means providing a maintenance strategy over an extended period is a major activity. Yet it also means that that minor enhancements to the maintenance regime can have major overall benefits in increasing efficiency and effectiveness of the service.

Historically many of the assets would have had their separate management regimes. This would be based on 'best' local management and operator experience. These approaches were generally revised with the formation of

the water companies in 1970s and initially the approach to assets renewal was ‘on the basis of nominal book lives’, but frequently ‘assets were replaced as they deteriorated’ but since ‘investment funds were limited to those directly affecting service … generally delayed beyond the book life’, Heather and Bridgeman (2006). Ansell et al (2011) illustrates the impact of delay in terms of cost to single items. The maintenance regime might take into account the knock-on effect on other assets within the surrounding network. Obviously such techniques would rarely look at the holistic impact across the whole maintenance strategy. So the ‘best’ local solution may not offer ‘best’ global solution and may penalise the whole organisation. Added to this there will be a set of constraints from the budget. Within the Water Industry this, may arise out of the “customers’ willingness to pay”. Some have questioned whether ‘willingness to pay’ is a sound constraint (Ugarelli et al, 2008), since again the customer may not be able to judge what will be ‘best’ in the long term. Within the UK the regulators will act as a partial safeguard for the Water Companies, since they supervise the level of service delivered and the price. There are also strong environmental constraints on Water Companies.

The Water Industry faces similar issues to other utilities. Therefore the models developed in the subsequent section are applicable across a wide range of industries beyond those in the utility industries. In this paper the models will be described, but for brevity only the base models will be discussed and not the full implementation. Hence a flavour of the approach will be given and not the full extension of the model. The focus will be on replacement of assets.

3. Models

The models described in this paper are being used within Scottish Water to aid strategic planning. Two types of models will be described. The first set of models is based on discrete time Markov Decision Process model (MDP). The second set is based on heuristics. Whilst the former has many benefits for a modelling view point, they do require certain assumptions to hold. Hence there are some situations where is it is necessary to consider other approaches to provide a solution.

In the description of the MDP approach to follow it will be assumed that the system is observed periodically with its state known at the beginning of a period. At this stage an action can be chosen from a set of known and allowable actions for that state. The result of the current state and the action means that an expected cost for the period will be incurred. The impact on the system during the period will mean that the state will evolve according to a known probability distribution which depends on the state and the action. Here the action set will be simply whether to replace or not. This can be formulated as a model as follows.

Assume the age of the current asset is i and the state of system is s, which comes from a set S, {0,1,2, …}. The decision (K_i = {0, 1}) is whether to keep (0) or replace (1) the current asset. The cost during the subsequent period is cost of general running (maintenance costs) and if appropriate a replacement

cost $c_i^0 = M_i$ and $c_i^1 = C_i + M_0$. As the state is defined to be age, the evolution of the process is deterministic in this case. When the decision to keep the asset is chosen, the asset will age by one during the subsequent period. When the decision to replace the asset is taken, the age is immediately reset to 0 and then the asset ages by one during the subsequent period. Hence the associated transition probabilities $p_i^0{}_{i+1}$ and $p_i^1{}_1$ are both 1.

In such a circumstance the long-run average cost would be obtained by minimising:

$$g + v_i = \min_{k \in K_i} (c_i^k + \sum_{j \in S} p_{i,j}^k v_j) \ \forall i \in S \quad (1)$$

where g is minimum average cost per period and $v_i - v_j$ is the relative advantage of being in state *j* rather than state *i.* This may not be regarded as a good solution since it is average over a long time horizon and costs are not going to remain the same. Hence it would be sensible to include a discount factor.

Assuming a discount factor on costs of β ($0 < \beta < 1$) then model might solution might be written as

$$v_i = \min_{k \in K_i} (c_i^k + \beta \sum_{j \in S} p_{i,j}^k v_j) \ \forall i \in S \quad (2)$$

Since it is unlikely that we would wish to run the problem over an infinite period it might be better to produce a finite horizon total reward model:

$$v_i^n = \min_{k \in K_i} (c_i^k + \sum_{j \in S} p_{i,j}^k v_i^{n-1}) \ \forall i \in S \quad (3)$$

where $v^N{}_i$ is the minimum expected cost for running the asset for N periods starting from state i. The aim is to find $v^N{}_i$ given values of $v^0{}_i$ where $v^n{}_i$ = minimum expected total cost over n periods starting from state i.

Again it still might be sensible to include a discount factor, into such a model and so one obtains the finite horizon discounted reward model:

$$v_i^n = \min_{k \in K_i} (c_i^k + \beta \sum_{j \in S} p_{i,j}^k v_i^{n-1}) \ \forall i \in S \quad (5)$$

This model can be extended to cover repair, refurbishment and replacement see Archibald et al (2004a, b).

Obviously so far the paper has only considered single assets and not multiple assets. It would be possible to simply use single asset results as a starting point to develop model for multiple assets. If the aggregation of single asset maintenance plans satisfies the constraints within the situation then it may be

acceptable as a strategy. It is, though, very likely that cumulative single asset plan will not satisfy the constraints (e.g. budget) within the situation. In such cases one might decide on a prioritisation amongst the assets. This may be to delay the replacement for some of the assets. This impact could be assessed using the insights gained from Ansell et al (2011).

As an alternative one might consider the 'restless bandits' approach (Whittle 1988) as discussed by Glazebrook et al (2002, 2006a and 2006b). In these latter papers Glazabrook at al suggested the use of the Whittle Index Heuristic as a solution.

Again assuming a discrete time model with an infinite horizon with discount factor β ($0 < \beta < 1$), but now assume there are N assets. Also only the decisions to keep or replace will be considered. Extensions are possible to cover repair, refurbishment and replacement but will be detailed subsequently. Each of the assets deteriorates independently according to a Markov process with the probability of transition from state i to state j for asset n given by $q_{i,j}(n)$. It will be assumed that in each period exactly L of the N items are selected to be replaced. The state of the assets can be described by $\boldsymbol{i} = (i_1, ..., i_N)$ where $S(n)$ is the set of possible conditions for item .Hence the state space for all the N assets can be described as $S = \{\boldsymbol{i}: i_n \in S(n)\}$.

As stated before the decision (k_i) is whether to keep (0) or replace (1) each asset i, so the set of decisions can be represented as $\boldsymbol{k} = (k_1, ..., k_N)$ and the action space is $K = \{\boldsymbol{k}: k_n \in \{0,1\} \ \& \ \sum k_n = L\}$ It is assumed that in any period there will be a maintenance cost and if applicable a replacement cost and so the cost will be $c_i^0(n) = M_i(n)$ & $c_i^1(n) = C_i(n) + M_0(n)$ for asset n in state i. The replacement and one period maintenance costs ($C_i(n)$ and $M_i(n)$ respectively for item n in state i) for each asset are non-decreasing with state. It is assumed that if replacement takes place then the state of system will be reset, and then each item deteriorates according to Markov chain $p_{\boldsymbol{i}\boldsymbol{j}}^{\boldsymbol{k}} = \prod_n q_{in(1-kn),jn}(n)$. The infinite horizon expected discounted cost is minimised by solving the following:

$$v_{\mathbf{i}} = \min_{\mathbf{k} \in K} \left(\sum_{n=1}^{N} c_{i_n}^{k_n}(n) + \beta \sum_{\mathbf{j} \in S} p_{\mathbf{i},\mathbf{j}}^{\mathbf{k}} v_{\mathbf{j}} \right) \ \forall \mathbf{i} \in S \tag{5}$$

where v_i = minimum infinite horizon expected discounted cost starting from state $\boldsymbol{i}$.

A difficulty arises with solving this problem, which is referred to as the curse of dimensionality. There are far too many possible states and this will often lead to the problem requiring impractical computational effort to solve. Whittle (1988), however, suggests a solution by suggesting there is a charge W for choosing replacement. This allows the solution to become more manageable under specific conditions.

Consider again the problem for a single asset and drop the dependency of costs and probabilities for convenience. Define $c_i^0 = M_i$ & $c_i^1 = C_i + M_0$. The

transition probabilities again assume deterioration according to Markov chain $p_i^0{}_j = q_{i,j}$ and $p_i^1{}_j = q_{0,j}$ respectively for keeping and replacement if applicable. The infinite horizon expected discounted cost is minimised by solving:

$$v_i = \min(c_i^0 + \beta \sum_{j \in S} q_{i,j} v_j, W + c_i^1 + \beta \sum_{j \in S} q_{0,j} v_j) \ \forall i \in S \qquad (7)$$

where v_i = minimum infinite horizon expected discounted cost starting from state *i.*

This will then provide a solution under the following situation. Let $\prod(W)$ be the set of states for which keeping is optimal. An item is said to be indexable if $\prod(W)$ is increasing in *W.* If an item is indexable, its Whittle index when in state i is the smallest *W* such that $i \in \prod(W)$. The Whittle index heuristic for the multiple asset problem replaces the *L* items with lowest Whittle indices (assuming positive) and keeps the rest.

The following is a simple illustration of the approach. Let the state of the asset be defined by the age of the asset. Let the costs be $c_i^0 = M_i$ and $c_i^1 = C + M_0$ and assume that the transition probabilities are simply $p_i^0{}_{i+1} = 1$ and $p_i^1{}_1 = 1$. Then it is possible to show that the Whittle index of item in state *i* is equal to

$$\frac{1-\beta^i}{1-\beta} M_i - C - \sum_{j=0}^{i-1} \beta^j M_j \qquad (8)$$

Obviously this approach has made the assumption that L assets will be replaced. More often the constraint on the system is in terms of the budget. Hence the total number of assets replaced is not the constraint, but rather the total spent. The question then becomes one of whether replacing many cheaper assets is more cost effective than replacing a single expensive asset. The Whittle Index, as described above, cannot answer this question, as it compares one asset to another, not one asset to many others. This can be resolved by normalising the model by dividing by the replacement cost of each asset. The resulting quantity is dimensionless, is 1 for a brand-new asset, and 0 for an asset at its optimal replacement age (or state), and can be viewed as a sort of 'fraction remaining value' for the asset. All assets will thus be treated similarly, and comparison between large and small assets becomes possible. A simple adaptation of the algorithm as illustrated in Figure 1 can hence ensure satisfaction of the budgetary constraint. This is the Normalised Whittle Index Heuristic.

Obviously there are a number of constraints on the above approach. It still may produce reasonable results even if the assumptions do not hold such as weakening the MDP requirement. For example it is possible to use search procedures such as a 'greedy solver' which iteratively look for the next best option one asset at a time. The advantage is that few assumptions are made. So far this has been implemented in an unconstrained form. A simple heuristic for the constrained case is to look for the best option for one year at a time which meets the constraints. This has been referred to as "Bang-per-buck"

strategy. Both these methods depend on the planning window or require the remaining life benefit value to be known.

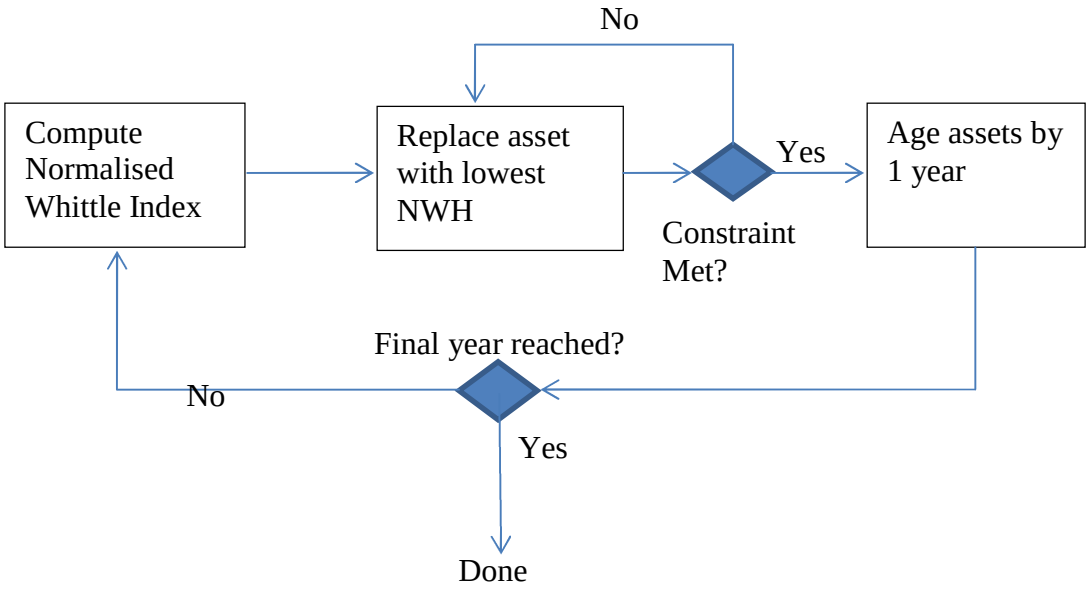


Figure 1: Normalised Whittle Index Heuristic

4. Application

As an illustration of the approach consider the following situation. The failure distribution is exponential and may deteriorate with age. Running cost are a, repairs can be made at cost b and the replacement cost is c. The discount factor is β. Suppose initially that a = £50, b = £100 and c = £500 with β = 0.96. Also the failure mechanism has a mean expected number of failures of exp(0.1age-1.1). Then Table 1 represents the Whittle Index Heuristic and it indicates that replacement should occur at 12.

Age, i	Whittle Index	Age, i	Whittle Index	Age, i	Whittle Index
1	-496.50	11	-96.50	21	2066.17
2	-488.92	12	5.33	22	2489.74
3	-476.59	13	124.99	23	2970.73
4	-458.80	14	264.79	24	3515.97
5	-434.69	15	427.31	25	4133.03
6	-403.34	16	615.43	26	4830.36
7	-363.71	17	832.36	27	5617.34
8	-314.60	18	1081.67	28	6504.38
9	-254.71	19	1367.36	29	7503.07
10	-182.56	20	1693.88	30	8626.28

Table 1: Example Whittle Index Heuristic

The Whittle Index would be repeated for the remainder of the assets and this would produce a list of replacement times and with each a Whittle Index value. Given that the model is set up initially to replace the first L items then these would be selected on the criteria of lowest costs.

Obviously the approach can be extended to take into account relative costs by considering the normalised Whittle Index which means implementation of the normalising algorithm presented in Figure 1.

Table 2 shows an illustration of impact on replacement age as the parameters a, b, c and β are changed. As the discount range decreases replacement age increases. As the relative cost of replacement decreases then replacement age decreases. As relative cost of repair decreases then replacement age increases. All these results seem intuitive.

a	B	c	β	Replacement age
50	100	500	0.96	12
50	50	500	0.96	16
50	100	250	0.96	10
50	50	250	0.96	12
50	100	500	0.9	14
50	50	500	0.9	18
50	100	250	0.9	10
50	50	250	0.9	14
50	100	500	0.8	16
50	50	500	0.8	21
50	100	250	0.8	19
50	50	250	0.8	16

Table 2: Impact on Replacement Age as parameters change

5. Conclusion

The paper produced the normalised Whittle Index to decide on optimal replacement strategy for assets where there are multiple assets based on a Markov Decision Process. It can be developed to for repair, refurbishment and replacement maintenance strategies. It should produce major savings for Utility companies. It provides an illustration of its implementation along with indication of sensitivity of the model.

References

1. Ansell JI, Archibald TW, and Thomas, LC, 'The Elixir of Life: Using a maintenance, repair and replacement model with virtual and operating age in the water industry, *IMA J Man Maths*, 15,151-160, (2004).
2. Archibald, TW, Ansell, JI, and Thomas, LC, *The Stability of an optimal maintenance strategy for repairable assets, J Process Mech Eng,* 218, 77-82, (2004).
3. Ansell, JI, Archibald, TW, Denning, R, and Bain, A, 'Investigating deferment of maintenance actions', *ARRTS Conference*, (April 2011).

4. Heather, AIJ; Bridgeman, ‘Water industry asset management: a proposed service-performance model for investment’, *Water and Environmental Journal*, Vol 21, 2, 127-132, (2007).
5. Ugarelli, R; Venkatesh, G; Brattebo, H; et al., ‘Importance of investment decisions and rehabilitation approaches in an ageing wastewater pipeline network. A case study of Oslo (Norway)’, *Water Science and Technology*, 58 (12), 2279-2293, (2008).
6. Glazebrook, KD, Nino-Mora J, and Ansell. PS, Index policies for a class of discounted restless bandit problems. *Advances in Applied Probability*, 34:754–774, (2002).
7. Glazebrook, KD, Ruiz-Hernandez, D, and Kirkbride, C, ‘Spinning plates and squad systems: Policies for bi-directional restless bandits’, *Advances in Applied Probability*, 38, 95-115, (2006a).
8. Glazebrook, KD, Ruiz-Hernandez, D, and Kirkbride, C, ‘Some indexable families of restless bandit problems’, *Advances in Applied Probability*, 38, 643-672, (2006b).
9. Whittle, P, Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, A25:287–298, (1988).

Road Network Flow Modelling for Maintenance

Chao Yang, Rasa Remenyte-Prescott and John Andrews
Nottingham Transportation Engineering Centre, University of Nottingham, Nottingham, NG7 2RD, UK

Abstract

Road works usually restrict the flow rate of vehicles on the network and affect network performance in terms of increased user costs and vehicle journey time. In addition to maintenance expenditure needed to keep roads at the required standard, the travel delay cost to road users caused by maintenance is significant and substantially exceeds the corresponding cost of maintenance. Therefore, in the highway asset management field there is a need to predict traffic congestion, so that road works can be planned considering their effects on the network.

In this paper a network level traffic flow model is presented, which is applicable to both motorway and urban road networks. A road network is modelled as a directed graph, containing roads and junctions, with different input flows throughout a day. The modelling is based on a simple idea of balancing out the traffic flow in the network, considering the capacity of an individual road and traffic overflow through to the related junctions. It is used to forecast travel delays and evaluate the network performance in normal conditions and when maintenance works are implemented. The model is used to illustrate how to calculate travel delays, how to estimate the effects of road works on the network and how to use those for maintenance planning.

1. Introduction

In recent years highways agencies have been moving their focus from construction of new roads to maintenance and rehabilitation of existing ones. Billions every year are spent to maintain UK roads at the required standard (AIA, 2011). In addition to maintenance cost, the travel delay cost due to maintenance is also significant and can substantially exceed the corresponding cost of maintenance. Therefore, it is of importance to consider road user costs in maintenance planning. The road network flow model, presented in this paper, is used to describe travel delays, when traffic flows and queues are predicted in a road network. Due to roadworks flow rates are restricted and queues on the network occur. In addition to modelling the network performance when no maintenance is performed, different maintenance strategies can also be compared in terms of travel delays in the network.

Typically, there are two main groups of traffic flow models used – macroscopic and microscopic. Macroscopic models are used to model the aggregate behaviour of sets of vehicles, whereas microscopic models are applied to the travel behaviour of an individual vehicle. Driver behaviour in real traffic is difficult to observe and measure and computational effort required for microscopic models can be high. A macroscopic model,

developed in this paper, is more suitable for modelling the traffic flow at the network level due to good real-time quality, realistic network representation and data availability.

A number of different macroscopic models have been constructed to model traffic on motorways and urban roads. For example, Lighthill and Whitham in 1955 developed a pioneering flow-dynamic model to describe unidirectional traffic flow on roads. In 1994 Daganzo proposed a cell transmission model that approximated the previous model to evaluate the network with a single entry and exit. METANET model, published by Messmer and Papageorgiou in 1990, considered a motorway network using such a model, and extensive research has been performed to improve it by introducing variable speed limits (Breton et al. 2002, Hegyi et al. 2005) and route guidance (Deflorio 2003, Karimi et al. 2004). In 2007 Van den Berg et al. developed a model for mixed urban and motorway networks, as motorway traffic is heavily influenced by the traffic flows on the connected urban roads, and vice versa. However, some properties of a road network structure could not be modelled using this approach. For example, priority junctions, i.e. T-junction and roundabout, were not considered; therefore, conflicting requirements of the traffic flows from competing directions could not be captured. Also, in the urban traffic model sub-queues for each direction of turning were assigned; thus, shared lanes where traffic heading to different directions might be mixed together were not taken into account. In addition to addressing these drawbacks, this paper evaluates both motorway and urban junctions based on the principle of a maximum capacity flow rate at the junctions where flows compete. Moreover, instead of applying an origin-destination trip matrix this model takes advantage of the observed junction turning ratios. Another feature is that in addition to some previous research on unidirectional traffic flow models, two-way traffic flow along network links is investigated in this model formulation.

The road network flow model is presented in this paper. The main steps of the method are given in Section 2, together with the illustration of the detailed model for two selected junction types. The case study on the local Loughborough-Nottingham network is presented in Section 3 in order to show the applicability of the model to calculate traffic delays on a simple network. Section 4 contains the overview of the model application for maintenance planning.

2. Road Network Flow Model

2.1 Network representation

In the proposed model the road network contains nodes and links. Nodes represent junctions, including signalized intersections and T-junctions, and links represent roads, including motorways and urban roads. Junction models have links which enable the exit traffic from one junction to enter the next junction and the process works in two directions to represent the two-way traffic. In this way, all the junctions are linked to each other. According to the network flow theory there can be a number of source and sink nodes on a network, where a source node is the flow into network and a sink node is the

flow out of the network. They are used to model the edges of the network. In addition, the links themselves can have source and sink nodes, which model cumulative traffic entering/leaving the link. This can represent significant traffic flows from/to such elements as housing estates, places of employment, airports and railway stations. Using this feature of the model it is possible to avoid the inclusion of minor roads on the network.

2.2 The main principle

The model is based on the idea of calculating queues on the network when due to the number of cars the flow capacity on the link is exceeded. First of all, the flow on each link in the network is calculated by passing flows from the source nodes to the sink nodes. Then the flow is compared to the flow capacity of each link and the queue is calculated if the flow capacity is exceeded, i.e. the rate of incoming vehicles is higher than the rate of outgoing vehicles. The queue is propagated back through the network if the link capacitance is exceeded, i.e. the queue is longer than the link. For every time step the same process is repeated and the effects are added to the queues already present in the network from the previous time steps. For example, due to the rush hour the flow on the link can increase and therefore, the queue on the link will also increase and even "overspill" to other incoming links. If the flow through the network improves, for example, traffic lights are adjusted to give a better link through the congested areas, the queues can decrease and eventually the network can become clear of queues. Following this principle the traffic on a given road network can be modelled throughout a day. Detailed rules are presented below in Section 2.2.

2.1 Notation

In the detailed description of the model the following notation is used:

$c_{i,j}$ flow capacity on the link between nodes *i* and *j* (passenger car unit), (pcu/hour)

$cp_{i,j}$ link capacitance, i.e. the maximum number of cars which can queue on the link from node *j* (pcu), which is obtained based on the length of the road from junction *i* to junction *j*

$sr_{i,j}(t_k)$ source flow entering the link at time t_k, for example, cumulative traffic joining the main road from an estate, (pcu/hour)

$sk_{i,j}(t_k)$ sink flow leaving the link at time t_k, for example, cumulative traffic leaving the main road for an estate, (pcu/hour)

$d_{i,j,l}(t_k)$ proportion of flow on the link choosing the outflow direction *l* at time t_k, *l* is expressed as the id of the destined node, for example, *j*+1

$f_{i,j}(t_k)$ flow on the link at time t_k, (pcu/hour)

$q_{i,j}(t_k)$ average number of vehicles queuing on the link at time t_k, (pcu)

$q_i(t_k)$ average number of vehicles propagating back to the upstream links of node *i* at time t_k, (pcu)

Δt length of time step, (hour)

2.2 Mathematical Model

Flow on the link *ij* at time t_k is calculated as a sum of all the flows to node *i*, the source flow entering the link and the negative sink flow leaving the link:

$$f_{i,j}(t_k) = \sum_{all\,s} d_{s,i,j}(t_k) f_{s,i}(t_k) + sr_{i,j}(t_k) - sk_{i,j}(t_k) \quad (1)$$

Once the flow on each link at time t_k is calculated, the flow and the queue values are updated according to the flow capacity $c_{i,j}$ and the link capacitance $cp_{i,j}$, if necessary. The updated flow and queue are expressed as $f'_{i,j}(t_k)$ and $q'_{i,j}(t_k)$ respectively. Three cases are considered:

1. Flow on the link is higher than the flow capacity and there is no queue at time t_k:

$$\text{If } f_{i,j}(t_k) > c_{i,j} \text{ and } q_{i,j}(t_k) = 0, \text{ then } f'_{i,j}(t_k) = c_{i,j} \quad (2)$$

$$\text{and } q'_{i,j}(t_k) = \left(f_{i,j}(t_k) - c_{i,j}\right) \cdot \Delta t \quad (2^a)$$

$$\text{If } q'_{i,j}(t_k) > cp_{i,j}, \text{ then } q'_{i,j}(t_k) = cp_{i,j}, \text{ and} \quad (2^b)$$

$$q_i(t_k) = \left(f_{i,j}(t_k) - c_{i,j}\right) \cdot \Delta t - cp_{i,j}$$

2. Flow on the link is higher than the flow capacity and there is a queue at time t_k:

$$\text{If } f_{i,j}(t_k) > c_{i,j} \text{ and } q_{i,j}(t_k) > 0, \text{ then } f'_{i,j}(t_k) = c_{i,j} \quad (3)$$

$$\text{and } q'_{i,j}(t_k) = q_{i,j}(t_k) + \left(f_{i,j}(t_k) - c_{i,j}\right) \cdot \Delta t \quad (3^a)$$

$$\text{If } q'_{i,j}(t_k) > cp_{i,j}, \text{ then } q'_{i,j}(t_k) = cp_{i,j} \text{ and} \quad (3^b)$$

$$q_i(t_k) = q_{i,j}(t_k) + \left(f_{i,j}(t_k) - c_{i,j}\right) \cdot \Delta t - cp_{i,j}$$

3. Flow on the link is lower than the flow capacity and there is a queue at time t_k:

$$\text{If } f_{i,j}(t_k) \le c_{i,j} \text{ and } q_{i,j}(t_k) > 0, \text{ then } f'_{i,j}(t_k) = c_{i,j} \quad (4)$$

$$\text{and } q'_{i,j}(t_k) = q_{i,j}(t_k) + \left(f_{i,j}(t_k) - c_{i,j}\right) \cdot \Delta t \quad (4^a)$$

$$\text{if } q'_{i,j}(t_k) < 0, \text{ then } f'_{i,j}(t_k) = c_{i,j} + \frac{q'_{i,j}(t_k)}{\Delta t} \text{ and } q'_{i,j}(t_k) = 0 \quad (4^b)$$

Note that once the queue is larger than the capacitance (Equations 2^b and 3^b), the queue at the end of the link, $q_i(t_k)$, is passed back to the network, i.e. to the links that contributed to the build up of the queue. The process of the queue propagation is simply described as passing back the proportion of the queue to each link, which equals the proportion of the flow that contributed to the overall flow in the first place. For example, if a queue builds up on the link between nodes *j* and *j*+1 and it exceeds the capacitance of the link by the number of vehicles, $q_j(t_k)$, it is proportionally distributed back to all the links that enter node *j*. This process increases the size of the queue and decreases the flow on each link that enters node *j*, as shown below:

$$q'_{i,j}(t_k) = q_{i,j}(t_k) + \frac{f_{i,j}(t_k)}{f_{j,j+1}(t_k)} \cdot q_j(t_k) \quad (5)$$

$$f'_{i,j}(t_k) = f_{i,j}(t_k) - \frac{f_{i,j}(t_k)}{f_{j,j+1}(t_k)} \cdot \frac{q_j(t_k)}{\Delta t} \quad (6)$$

The process of queue propagation is carried out through the network until a queue can be accommodated and does not exceed the capacity of the link. If the queue is present at time t_k, it is also present at the beginning of the next time step t_{k+1}, i.e. $q_{i,j}(t_{k+1}) = q'_{i,j}(t_k)$. Note that nodes on the edges of the network are assigned to infinite flow capacity so that the process of queue propagation is terminated at the boundaries of the network.

2.3 Junction Models

The sub-models for each junction type are constructed to express the traffic interaction at junctions. The junction types considered in this study are listed in Table 1.

Junction groups	Junction types
Signalised junctions	Signalised T-junction, signalised intersection, signalised roundabout
Priority Junctions	T-junction, urban roundabout, motorway roundabout
One-way junctions	On-ramp and off-ramp, merge and diverge, roadwork node

Table 1. Junction types

Some less common junctions are explained below. For example, on the motorway roundabout vehicles can queue on the roundabout itself when the exit to the motorway is blocked and vehicles are queuing to join the motorway. On-ramp and off-ramp junctions are used to model slip roads for entering and exiting the motorway. A roadwork node represents a part of the road under maintenance.

To overview the main features of the models, the traffic flow at a signalised junction is influenced not only by the flow capacity of the entry links, but also by the green split time of the signals, i.e. the proportion of times the signals give the priority to the flow in this direction. For priority junctions the traffic flow is based on the right-of-way rules, when the entering flow is restricted not only by the flow capacity but also by the traffic flows from the competing links. All these features have been considered in the model, and to demonstrate the concept the T-junction model is described below.

T-junction model: As shown in Figure 1 such a junction is controlled by the right-of-way rules, i.e. a vehicle travelling on the major road has the priority

and a vehicle approaching the major road must allow the traffic to pass before joining the major road.

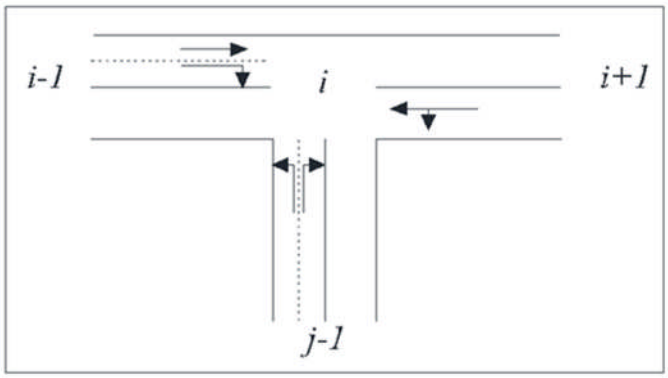


Figure 1. T-junction model

It is assumed that some roads have a single lane, for example, the right part of the major road, and some roads have two lanes, for example, the left part of the major road, where lane 1 is used for going straight on and lane 2 is used for turning right and crossing the oncoming traffic. However, a different set-up of lanes can be implemented in the model as necessary. Before the mathematical model for the T-junction is presented, additional notation is introduced:

$f_{i-1,i,l}(t_k)$	flow at node *i* that coming from direction *i*-1 and going in lane *l*, i.e. 1 represents the left lane, at time t_k, (pcu/hour)
$c_{i-1,i,l}$	node *i* flow capacity for the traffic coming from direction *i*-1 and going in lane *l*, depending on gaps between vehicles and vehicle speed, (pcu/hour)
$cp_{i-1,i,l}$	link capacitance in lane *l*, i.e. the maximum number of cars which can queue in lane *l* of the link, (pcu)
$q_{i-1,i,l}(t_k)$	average number of vehicles queuing on lane *l* of road *i*-1 for node *i* at the beginning of t_k, (pcu)

In the model the three roads on the T-junction are considered one by one:

I. The major road *i*-1

For the flow from *i*-1 to *i*+1, i.e. in lane 1, no restrictions on the flow from conflicting traffic exists. Therefore, a queue can only build up due to restrictions upstream the network and the general rule to calculate it, described in Section 2.2, is used.

For the flow from *i*-1 to *j*-1, i.e. in lane 2, the conflicting flow is the flow from *i*+1 to *i*. A queue builds up if the flow in lane 2 or the conflicting flow is higher than the flow capacity in lane 2 through the intersection. Five cases are considered, according to different relationships between the flow, its conflicting flow, the capacity and the capacitance of the relevant link:

1. Assume that the flow in lane 2 is higher than the flow capacity in lane 2 through the intersection, the conflicting flow from direction *i*+1 is lower than the flow capacity in lane 2 through the intersection, and there is no queue at time t_k:

If $f_{i-1,i,2}(t_k) > c_{i-1,i,2}$, and $f_{i+1,i}(t_k) \le c_{i-1,i,2}$ and $q_{i-1,i,2}(t_k) = 0$, then (7)

$$f'_{i-1,i,2}(t_k) = c_{i-1,i,2} \text{ and } q'_{i-1,i,2}(t_k) = \left(f_{i-1,i,2}(t_k) - c_{i-1,i,2}\right) \cdot \Delta t \quad (7^a)$$

If $q'_{i-1,i,2}(t_k) > cp_{i-1,i,2}$, then $q'_{i-1,i,2}(t_k) = cp_{i-1,i,2}$ and (7^b)

$$q_{i-1}(t_k) = \left(f_{i-1,i,2}(t_k) - c_{i-1,i,2}\right) \cdot \Delta t - cp_{i-1,i,2}$$

2. The conflicting flow from direction *i*+1 is higher than the flow capacity in lane 2 through the intersection and there is no queue at time t_k:

If $f_{i+1,i}(t_k) > c_{i-1,i,2}$ and $q_{i-1,i,2}(t_k) = 0$, then $f'_{i-1,i,2}(t_k) = 0$ and (8)

$$q'_{i-1,i,2}(t_k) = f_{i-1,i,2}(t_k) \cdot \Delta t \quad (8^a)$$

If $q'_{i-1,i,2}(t_k) > cp_{i-1,i,2}$, then $q'_{i-1,i,2}(t_k) = cp_{i-1,i,2}$, and (8^b)

$$q_{i-1}(t_k) = f_{i-1,i,2}(t_k) \cdot \Delta t - cp_{i-1,i,2}$$

3. Flow on the link in lane 2 is higher than the flow capacity in lane 2 through the intersection, the conflicting flow from direction *i*+1 is lower than the flow capacity in lane 2 through the intersection and there is a queue at time t_k:

If $f_{i-1,i,2}(t_k) > c_{i-1,i,2}$, and $f_{i+1,i}(t_k) \le c_{i-1,i,2}$ and $q_{i-1,i,2}(t_k) > 0$, then (9)

$$f'_{i-1,i,2}(t_k) = c_{i-1,i,2} \text{ and } q'_{i-1,i,2}(t_k) = q_{i-1,i,2}(t_k) + \left(f_{i-1,i,2}(t_k) - c_{i-1,i,2}\right) \cdot \Delta t \quad (9^a)$$

If $q'_{i-1,i,2}(t_k) > cp_{i-1,i,2}$, then $q'_{i-1,i,2}(t_k) = cp_{i-1,i,2}$, and (9^b)

$$q_{i-1}(t_k) = q_{i-1,i,2}(t_k) + \left(f_{i-1,i,2}(t_k) - c_{i-1,i,2}\right) \cdot \Delta t - cp_{i-1,i,2}$$

4. The conflicting flow from direction *i*+1 is higher than the flow capacity in lane 2 through the intersection and there is a queue at time t_k:

If $f_{i+1,i}(t_k) > c_{i-1,i,2}$ and $q_{i-1,i,2}(t_k) > 0$, then $f'_{i-1,i,2}(t_k) = 0$, and (10)

$$q'_{i-1,i,2}(t_k) = q_{i-1,i,2}(t_k) + f_{i-1,i,2}(t_k) \cdot \Delta t \quad (10^a)$$

If $q'_{i-1,i,2}(t_k) > cp_{i-1,i,2}$, then $q'_{i-1,i,2}(t_k) = cp_{i-1,i,2}$, and (10^b)

$$q_{i-1}(t_k) = q_{i-1,i,2}(t_k) + f_{i-1,i,2}(t_k) \cdot \Delta t - cp_{i-1,i,2}$$

5. Flow on the link in lane 2 is lower than the flow capacity in lane 2 through the intersection, the conflicting flow from direction *i*+1 is lower than the flow capacity in lane 2 through the intersection and there is a queue at time t_k:

If $f_{i-1,i,2}(t_k) \le c_{i-1,i,2}$, and $f_{i+1,i}(t_k) \le c_{i-1,i,2}$ and $q_{j-1,i,l}(t_k) > 0$, then (11)

$$f'_{i-1,i,2}(t_k) = c_{i-1,i,2} \text{ and } q'_{i-1,i,2}(t_k) = q_{i-1,i,2}(t_k) + \left(f_{i-1,i,2}(t_k) - c_{i-1,i,2}\right) \cdot \Delta t \quad (11^a)$$

$$\text{If } q'_{i-1,i,2}(t_k) < 0, \text{ then } f'_{i-1,i,2}(t_k) = c_{i-1,i,2} + \frac{q'_{i-1,i,2}(t_k)}{\Delta t} \text{ and } q'_{i-1,i,2}(t_k) = 0 \quad (11^b)$$

II. The major road *i*+1

For the flow from direction *i*+1 no restrictions from the conflicting traffic exist. Therefore, a queue can only build up due to restrictions upstream the network and the general rule, described in Section 2.2, is used.

III. The minor road *j*-1

For the flow from *j*-1 in lane 1, the conflicting flow is the flow from *i*+1 to *i*-1, while for the flow in lane 2, the conflicting flow is the sum of the flow from *i*+1 to *i*-1 and the flow from *i*-1 to *i*+1. Both flows on this road are evaluated as the flow in lane 2 from direction *i*-1, described above.

Roadwork node: Such nodes have been introduced for investigation of maintenance effects on traffic delays. Three types of roadwork nodes were considered, i.e. roadwork nodes used to model maintenance on a single carriageway, on a dual carriage way and on a motorway. For example, it is assumed that on a single carriageway, two opposing traffic flows have to share the single lane in service, controlled by traffic lights, when maintenance is implemented. Therefore, the traffic can pass the worksite through the corresponding green splits, as for the rest of time the traffic has to wait and leave gaps for the opposing traffic. The model for such roadwork arrangements includes not only the length of the closed part of the link but also the duration of green splits. For example, on a dual carriageway or a motorway one or two lanes can be closed for maintenance and speed restrictions are usually applied on the lane in service. Therefore, the model includes the length and the width of the closed part and the reduced flow capacity in the lane passing the roadworks.

The application of the method including the most common junction types and roadwork nodes is presented in Section 3.

2.4 Network Performance Metrics

After calculating queues on the individual links, the performance metrics for a network is derived as the aggregation of the performance for all the junctions and all the links, expressed in travel time and road user cost. The total travel duration spent in the network at time step t_k is defined as:

$$TTD(t_k) = \sum_{i=1}^{N} T_{D,i}(t_k) + \sum_{j=1}^{M} T_{T,j}(t_k) \tag{12}$$

$$T_{D,i}(t_k) = \sum_{a=1}^{A_i} \int_{t_{k-1}}^{t_k} q_{a,i}(t)dt \tag{13}$$

where

$T_{D,i}(t_k)$ delay time spent by road users at junction *i* at time t_k, (h)

$q_{a,i}(t_k)$ length of queue on arm *a* to junction *i* at time t_k, (pcu)

A_i number of arms at junction *i*

N number of junctions on the network

$T_{T,j}(t_k)$ journey time spent on link j, (h)

M number of links on the network

Note that junction arms are the links entering and exiting the junction. Also, for each link on the network the vehicle journey time, $T_{T,j}(t_k)$, is calculated considering the following values of time as appropriate at each time step: the travel time for the non-disturbed traffic that pass through the link without delays, the time spent on queue formation and the time spent on queue dissipation.

Finally, the total user cost spent on the network is evaluated as:

$$C_u = TTD(t_k) \times v \tag{14}$$

where

v time value of road user; according to 2012 prices and values the market price value of time for an average vehicle is £15.38 per hour (DfT, 2012)

2.5 Model Data

For the modelling of travel delays using this approach, geographical characteristics of the road network are needed along with the traffic flows on the network at different times of the day. In terms of geographical features, the length of each link and the flow capacity for each arm in the junction are required. In terms of traffic data, the traffic entering each link during the day, the proportion of vehicles leaving each junction on each exit arm and the signal control inputs are each time interval are needed.

2.6 Network Solution Routine

The whole set of traffic flows is interconnected, since the inflows at each junction come from the outflows of the upstream network and, at the same time, the outflows are functions of the inflows for the junction. Therefore, at the beginning of the process the sequence of nodes for evaluation is chosen. The simulation process starts at some original node, it progresses through the network and ends at the same node. Then the simulation is processed iteratively until the convergence is reached, i.e. the outflows cannot be changed by further iterations. Once the parameters reach a stable state the effects of vehicles that propagate back to the upstream links are calculated, which can make traffic delays even more severe. Since the results at the current time step become the input at the next time step, traffic delays over a time period can be evaluated.

3. Case Study

A road network in the area of Loughborough and Nottingham has been used to illustrate the applicability of the method to predict travel delays and estimate different maintenance approaches on a road network.

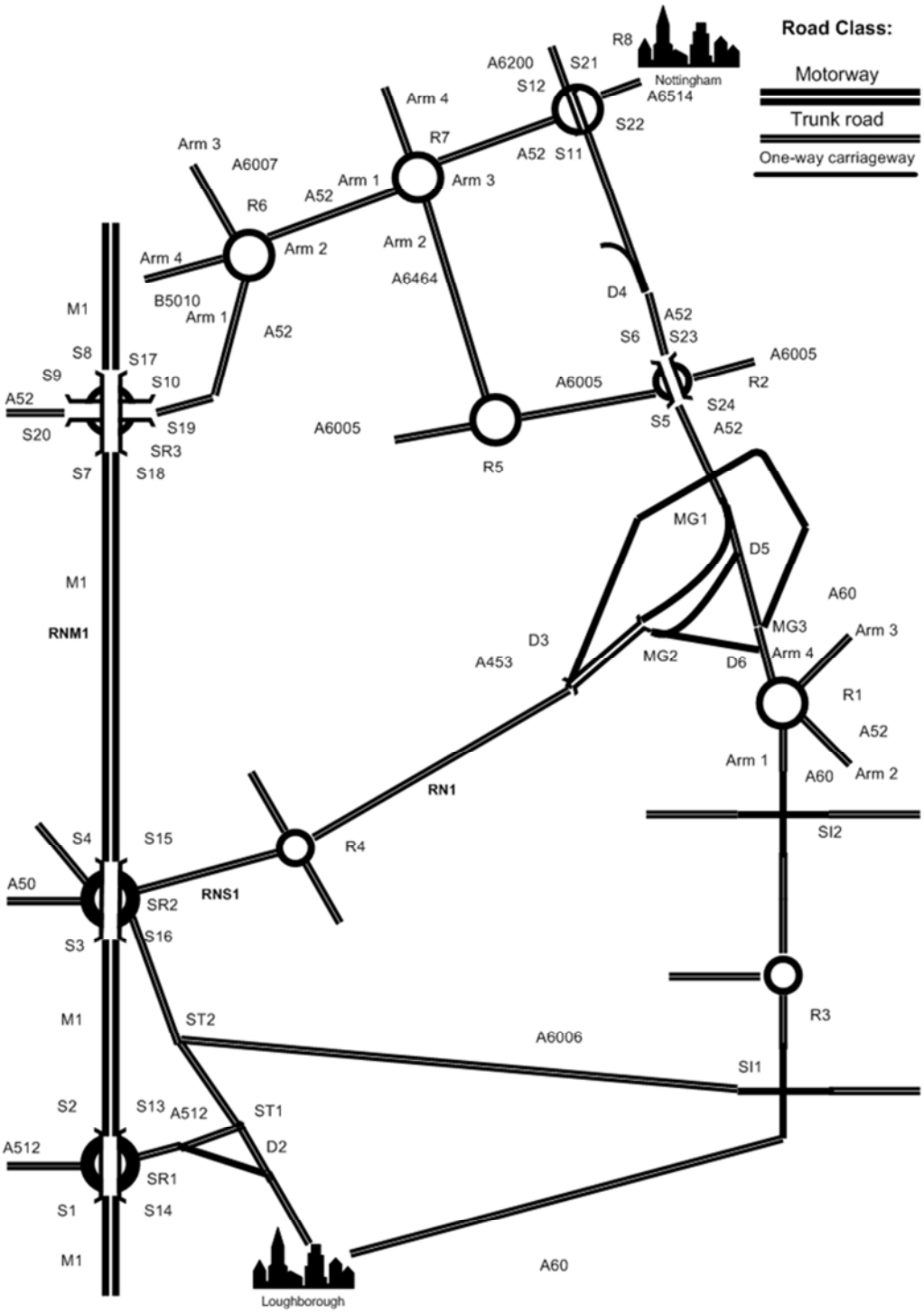


Figure 2. Loughborough-Nottingham example network

3.1 Example Network

The network, shown in Figure 2, is composed of trunk roads, regional rural and urban roads. For simplicity, only trunk roads and roads that connect to major junctions are retained in the network. Some symbols, e.g. M1 and A52, represent the road class and road number of the road links, while other symbols denote the type and id of the junctions in the model, for instance, the

diverge junction in Loughborough is named D2. There are 47 junctions modelled in the network, including 8 roundabouts (R), 3 signalised roundabouts (SR), 5 diverge junctions (D), 3 merge junctions (MG), 12 off-ramps and 12 on-ramps (both denoted by S), 2 signalised T-junctions (ST) and 2 signalised intersections (SI).

3.2 Model Data

The available data for the network was a set of traffic data which in the last few years had been collected at various locations on trunk roads and at various junctions, also containing inflows and proportions of turning traffic for most junctions. Such traffic data was obtained from the Highways Agency [31] and Nottingham County Council. The data was applied in the form of a two-way traffic flow every hour.

3.3 Example Network Performance

The objective of this case study was to test the method on a simple network, predict the outflows throughout the day and identify weak areas of the network which experience severe traffic congestion. For the illustration purposes in this paper, the algorithm was executed for 16 time steps on a working day, i.e. 16 hourly steps between 7am to 11pm. From the modelling results in Table 2, it was found that the roundabouts R1, R6 and R7 and the incoming links were badly congested during the morning and afternoon peak times, while the rest of the network could cope with the entering flows. Note that these three roundabouts are on the Nottingham ring road, they carry most of the traffic in the network and usually are badly congested at these times, as confirmed by the model.

Time	Traffic condition states				
	Arm 1 (R1)	SI2-R1	Arm 3 (R6)	Arm 4 (R6)	Arm 4 (R7)
7:00	NQ	NQ	NQ	NQ	NQ
8:00	MINQ	NQ	NQ	MAJQ	MAJQ
9:00	MAJQ	MINQ	NQ	MAJQ	NQ
10:00	MAJQ	MAJQ	MAJQ	NQ	NQ
11:00	NQ	NQ	NQ	NQ	NQ
12:00	NQ	NQ	NQ	NQ	NQ
13:00	NQ	NQ	NQ	NQ	NQ
14:00	NQ	NQ	NQ	NQ	NQ
15:00	NQ	NQ	NQ	NQ	NQ
16:00	NQ	NQ	NQ	NQ	NQ
17:00	NQ	NQ	NQ	MAJQ	MAJQ
18:00	NQ	NQ	NQ	MAJQ	MAJQ
19:00	NQ	NQ	MAJQ	NQ	NQ
20:00	NQ	NQ	NQ	NQ	NQ
21:00	NQ	NQ	NQ	NQ	NQ
22:00	NQ	NQ	NQ	NQ	NQ

Table 2. Traffic condition for the congested areas in the Loughborough-Nottingham network

Note that “NQ” represents that no queue formed in the link, “MINQ” represents that minor queue formed in the link (under the threshold of 100 vehicles), “MAJQ” represents major queue formed in the link (more than 100 vehicles).

The same conclusion can be drawn from the results in Figure 3, when all the queues on the network were added at each time step to represent the overall network performance.

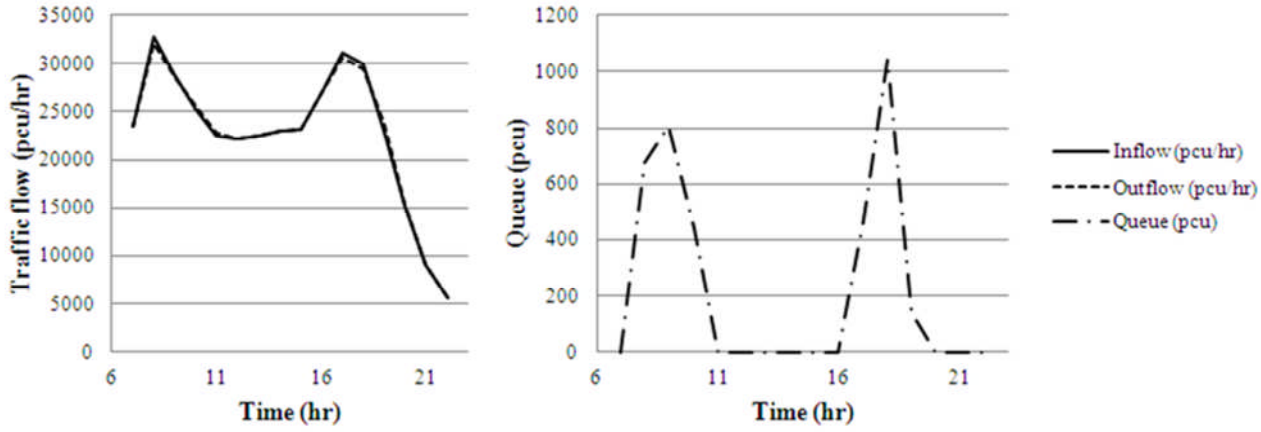


Figure 3. Traffic condition for the Loughborough-Nottingham network without maintenance

Once the congested areas on the network were found, the method has been further developed to reroute the traffic around the network, so that the traffic delays are minimised (*The work on rerouting has been completed by the authors but is not the scope of this paper*).

3.4 Maintenance Effects on Network Performance

Maintenance works were modelled by using the roadwork nodes as described in Section 2.3. To illustrate the concept, consider a road link SR2-R4 on the example network, which is a single carriageway, and the worksite on it, defined as RNS1 in Figure 2. The results on traffic congestion are given in Figure 4. For comparison, it captures three cases: in case 1 no maintenance is performed; in case 2 the green splits for both directions are assumed as 45% and in case 3 the length of worksite is the double of it in case 2, therefore, the green splits are assumed as 40%. The results show that the travel delay in the network during maintenance is much greater than in the network under normal conditions due to lane closure during maintenance. They also show that the queues in case 3 are longer than in case 2 due to the longer worksite and shorter time for the green splits.

The results in Figure 4 confirm that maintenance work has a great impact on the network performance, especially when the normal flow along a road is very much higher than the capacity, i.e. at morning and afternoon peaks. By analysing the travel delay incurred by maintenance, roadworks can be

planned to cause less disruption when possible, i.e. perform maintenance when the traffic is low. However, additional factors, such as the increased cost of maintenance at night, would need to be considered in maintenance planning.

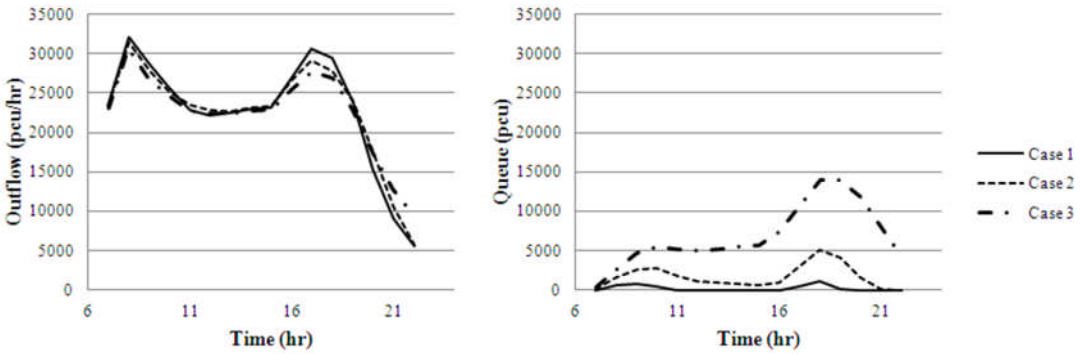


Figure 4. Traffic condition for the example network under maintenance

4. Model Application for Maintenance Planning

Further studies have been performed on how the road network flow model can be used to analyse effects on the network when different maintenance works are carried out. For example, different start times of maintenance during the day and different maintenance worksite arrangements, such as the number of closed lanes and traffic controls of flows, can be evaluated. In this case, the total cost is estimated including the highway agency cost and the road user cost, and a way of how to minimise the traffic congestion when maintenance is implemented can be proposed.

Consider an example network with its daily performance without maintenance shown in Table 3. The chosen network contains 6 junctions and 10 links and is used to illustrate the usage of the model to plan maintenance.

Time	Inflow (pcu/hr)	Outflow (pcu/hr)	Queue (pcu)	Journey time (h)	Travel delay (h)	Total travel duration (h)	Road user cost (£)
7-8	10000	10000	0	538	0	538	8267
9-15	5000	5000	0	269	0	269	4133
16-18	10000	10000	0	538	0	538	8267
19-21	5000	5000	0	269	0	269	4133
22	2000	2000	0	108	0	108	1653
23	1000	1000	0	54	0	54	827
0-5	100	100	0	5	0	5	83
6	1000	1000	0	54	0	54	827

Table 3. Illustration of daily network performance without maintenance

This road network is sufficient for delivering the required traffic flow and no traffic delays are present.

Assume that a single carriageway on this network needs to be closed for maintenance for 10 hours and the maintenance cost is £6250. The purpose of the analysis is to find the best start time of works during the day. It is assumed that maintenance should be completed within 24 hours and the start of the day is considered at 7am. Initially, maintenance works are considered to start at 9am, and the results are shown in Table 4.

Time	Inflow (pcu/hr)	Outflow (pcu/hr)	Queue (pcu)	Journey time (h)	Travel delay (h)	Total travel duration (h)	Road user cost (£)
7-8	10000	10000	0	538	0	538	8267
9-15	5000	5000	0	272	0	272	4191
16	10000	9624	376	520	189	708	10890
17	10000	9412	964	519	672	1191	18322
18	10000	9355	1609	527	1291	1817	27946
19	5000	6609	0	384	297	681	10476
20-21	5000	5000	0	269	0	269	4133
22	2000	2000	0	108	0	108	1653
23	1000	1000	0	54	0	54	827
0-5	100	100	0	5	0	5	83
6	1000	1000	0	54	0	54	827

Table 4. Illustration of maintenance effects, maintenance start time is 9am

The road user cost increased significantly and some queues were formed, since the opposing flows had to use the single available lane due to single carriageway maintenance. The total cost of this maintenance arrangement is £125572, obtained by summing the daily road user cost and the maintenance cost. The delays could be minimised by changing the start time of maintenance, as shown in Figure 5.

Figure 5. Illustration of obtaining the best start time

Figure 5 indicates that the maintenance scenario with start time at 9pm led to the least total cost, since only a small number of vehicles were entering the network at night.

In addition, using the analysis the worksite arrangements can be chosen in a way that minimises traffic delays. Figure 6 illustrates that the green splits can be optimised to reduce traffic delays, showing the performance of a network with different green splits during maintenance of a single carriageway. In this example, the best green splits are in the interval [40%, 50%].

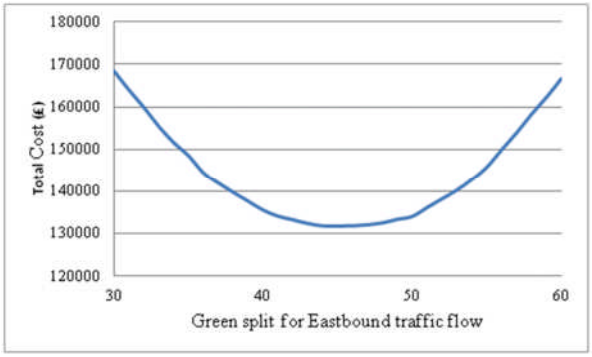


Figure 6. Illustration of obtaining the best green splits

Once maintenance effects on the network were modelled, optimisation techniques were applied to find best maintenance strategies for the network. The best solutions were searched for considering the two conflicting parties – road users and highway agency, since performing road maintenance at night can be convenient for road users but costly for highway agency. Short-term and long-term strategies were explored. (*The work on optimisation has been completed by the authors but is not the scope of this paper*).

5. Conclusions and Future Work

In this paper a macroscopic road network flow model has been proposed. The model calculates the flows through the network and the queues which build up and disperse at different points during the day. The modelling capability advances the previously developed macroscopic traffic flow models in the following features:

- It accounts for both motorway and urban roads in the same network and reflects the interactive nature of the two systems
- It models two-way traffic flow by using an iterative simulation method to calculate dependent traffic flows in the network
- It accounts for traffic exiting or joining the network along an urban network link to simulate the traffic from concentrated points, such as housing estates or work place locations
- It deploys shared lane to illustrate the traffic interaction among mixed directional traffic flows
- It models roadworks on the network and has the capability to consider the geometry and traffic control on a worksite.

The results presented indicate that the model is capable of describing the evolution of dependent traffic flows and forecast the traffic movement and queue dynamics through a road network. Also, the model is suitable to demonstrate maintenance effects on the traffic condition, when the network

undergoes road maintenance. Different maintenance arrangements can be evaluated in terms of congestion.

In the future in order to address potential processing time issues if larger road networks are modelled, the performance of the network solution routine could be improved by minimising the time spent to evaluate the network. Also, the model could be exploited for further applications in maintenance planning. For example, in addition to maintenance cost and road user cost considered, road condition data could be taken into account and different types of maintenance, suitable for a specific road condition, could be investigated.

ACKNOWLEDGMENT
John Andrews is the Royal Academy of Engineering and Network Rail Professor of Infrastructure Asset Management. He is also Director of The Lloyd's Register Foundation* Centre for Risk and Reliability Engineering at the University of Nottingham. Rasa Remenyte-Prescott is the Lloyd's Register Foundation Lecturer in Risk and Reliability Engineering. They would both like to express their gratitude to all of these organisations for their support.
* The Lloyd's Register Foundation (LRF) supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

References

1. A. I. A. (2011) Annual Local Authority Road Maintenance Survey.
2. Lighthill, M. J. and G. B. Whitham (1955). "On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads." Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **229**(1178): 317-345.
3. Daganzo, C. F. (1994). "The Cell Transmission Model - a Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory." Transportation Research Part B-Methodological **28**(4): 269-287.
4. Daganzo, C. F. (1995). "The Cell Transmission Model .2. Network Traffic." Transportation Research Part B-Methodological **29**(2): 79-93.
5. Messmer, A. and M. Papageorgiou (1990). "METANET: A macroscopic simulation program for motorway networks." Traffic Engineering and Control **31**(9): 466-470.
6. Breton, P., A. Hegyi, et al. (2002). Shock wave elimination/reduction by optimal coordination of variable speed limits. Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference.
7. Hegyi, A., S. Bart De, et al. (2005). "Optimal coordination of variable speed limits to suppress shock waves." Intelligent Transportation Systems, IEEE Transactions on **6**(1): 102-112.
8. Deflorio, F. P. (2003). "Evaluation of a reactive dynamic route guidance strategy." Transportation Research Part C: Emerging Technologies **11**(5): 375-388.
9. Karimi, A., A. Hegyi, et al. (2004). Integration of dynamic route guidance and freeway ramp metering using model predictive control. American Control Conference, 2004. Proceedings of the 2004.
10. Van den Berg, M., A. Hegyi, et al. (2007). "Integrated traffic control for mixed urban and freeway networks: A model predictive control approach."

European Journal of Transportation and Infrastructure Research **3**(3): 223-250.
11. HA (2011) Traffic Information Database HATRIS.
12. DfT (2011) Transport Analysis Guidance WebTAG, Section 3.5.6.

Probabilistic reliability and risk analysis for systems of fusion device

Roman Voronov, Robertas Alzbutas
Laboratory of Nuclear Installation Safety, Lithuanian Energy Institute,
3 Breslaujos str., LT-44403 Kaunas, Lithuania

Abstract

A number of fusion devices are under construction in Europe. Since fusion energy is innovative and fusion devices contain unique and expensive equipment, an issue of their reliability is very important from their efficiency perspective.
Reliability, Availability, Maintainability, Inspectability (RAMI) analysis is being performed or is going to be performed in near future for such fusion devices as ITER and DEMO in order to ensure reliable and efficient operation for experiments (ITER) or energy production (DEMO) purposes. On the other hand, rich experience of reliability and probabilistic safety analysis (PSA) exists in nuclear industry for fission power plants and other nuclear installations.
In this paper the Wendelstein 7-X (W7-X) device is mainly considered. It is a stellarator type fusion device under construction in the Max-Planck-Institut für Plasmaphysik, Greifswald, Germany (IPP) . In frame of cooperation between IPP and Lithuanian Energy Institute (LEI) under EURATOM treaty a pilot project of reliability analysis of the W7-X systems was performed with a purpose to adopt NPP PSA experience for fusion device systems. During the project a reliability analysis of a divertor target cooling circuit, which is an important system for permanent and reliable operation of in-vessel components of the W7-X was performed.

1 Introduction

In general, the purpose of reliability and risk analysis is to provide support in making correct management decisions by evaluating the reliability and risk associated with a set of decision alternatives. Classical definition of the risk of failure is:

$$R=p_f C; \tag{1}$$

where R is the risk of failure, p_f is a kind of a measure of reliability, i.e. the probability of failure and C is the cost caused by the failure. The measure of the failure cost may be different (depending on various consequences). For production plants (including nuclear power plant) it is usually not only the cost of failure (and accident in the worst case) and repair itself but also an amount of lost production (e.g. electricity) and lost profit.
Risk can be reduced from a level R to a lower level R' either by reducing the loss given failure or by reducing the probability of failure, or even by reducing both parts [1]. From the other hand, such risk reduction requires some investments and should be taken into account during the analysis. The values Δpf and ΔC should be selected in such a way that the risk reduction ΔR is achieved at minimal cost.
The purpose of this paper is to demonstrate how the reliability, as the main ingredient of safety (antonym of the risk), could be analysed for systems of fusion devise and show the practical application and results of such analysis by possibility to reduce the risk and the cost in relation to the risk. In section 2, we will review the approaches of reliability analysis used for fusion devices; section 3 is devoted for short survey of methods and

techniques used for the analysis; in section 4 we will demonstrate the results of these methods practical application as a case study for Divertor Target Cooling Circuit ACK10 of Wendelstein 7-X experimental fusion device.

2 Reliability analysis of fusion devices

Power plant availability is essential from the economical perspective as both fission and fusion power plants require very high initial investments. Returning of the investment and earning profit requires the plant to generate the highest possible amount of electricity and this implies high availability requirement.
High Availability of experimental fusion devices is required for most efficient use of the device for experiments.
A conceptual study of future commercial fusion power plants (FPPs) has been performed with a Power Plant Availability (PPA) study aimed at identifying the aspects affecting the availability and generating costs of FPPs [2, 3]. Among others, availability and reliability issues of FPPs were covered by the study. The study concludes, that in order to be competitive, fusion plants starting from the first generation need to comply with availability factor greater than 80%, similar to existing fission plants, with very few unplanned shutdowns. In order to guarantee continued safety of operation during fusion plant plant lifetime, in-service inspection and maintenance are needed and this aspect should be taken into consideration in the design of the systems [2, 4].

2.1 ITER RAMI Approach

Availability objective for ITER is 60% inherent availability and 32% operational availability [3]. The inherent availability is the percentage of time during which the machine would be available if no delay due to scheduled maintenance or supply was encountered. The operational availability reflects the inherent design including the effects of of maintenance/upgrade delays taking into account the availability of maintenance personnel and spares and other non-design factors.
ITER organization uses RAMI (Reliability, Availability, Maintainability, Inspectability) approach to perform a technical risk assessment. RAMI approach focuses on the operational functions required by the operation of ITER rather than on physical components. It enables to define requirements for the operational functions and provide the means to ensure that they could be met.

The RAMI process begins during the design phase of a system because corrective actions are still possible. The process is focused on the functions required to operate ITER and their failure criticality. It is declined in 4 steps:

- Functional Analysis (FA);
- Failure Modes, Effects and Criticality Analysis (FMECA);
- Risk mitigation actions;
- RAMI requirements.

Functional analysis of the systems is performed with a functional breakdown (top-down description of the system as a hierarchy of functions) and an assessment of reliability and availability performance of the functions by using Reliability Block Diagrams (RBDs). The RBD approach uses the function blocks (FB) as a basis, but concentrates on the reliability-wise relationships between the function blocks. The input data, such as

mean time between failures (MTBF) and mean time to repair (MTTR), are fed to the lowest level blocks.
A FMECA is performed in parallel to the RBDs to list the function failure modes and evaluate their risk level. A decision whether to accept or mitigate the failure mode is made based on the risk level. FB and RBD are input to FMECA.
Risk mitigation actions are initiated in order to reduce the risk level of the failure modes identified by the FMECA. After integrating RMA, the new RBDs are prepared.

RAMI requirements are outputs of the ITER RAMI process. They are integrated in the system requirements:

- Availability and reliability targets for the system and main functions according to the project requirements.
- Required design changes that need to be integrated to improve the current design.
- Specific tests to be performed on the components or systems.
- Operation procedures and specific training to lower the risks when operating the machine.
- Maintenance requirements in terms of list of spares, intervals of inspection and preventive maintenance, procedures and training.
- Proposals for standardization of common parts used in great number in the project, as ensuring inter-changeability of spares in the design of the systems shall then allow for shorter maintenance operation (replacement of consumables, repairs of failed components) and shall reduce the downtime of the systems and the Severity ratings in the FMECA, reducing the risk level and allowing for more availability of ITER for the experimental programme).

The process applied for the analysis of the plant systems defines failures of the functions, their criticality and provides risk mitigation actions. Up to 2010 RAMI was applied to 16 out of 21 main ITER systems. Analysis performed for the Tokamak Cooling Water System [5] identified initially 27 major risks, such as failure of the main pumps, leaks on the heat exchangers or associated valves leading to loss of cooling and possible damage for the plasma-facing components and failure of the coolant chemistry control leading to corrosion. For such major risks risk mitigation actions are taken which reduce either the likelihood (prevention) or the consequences (protection) of the failures. Analysis proves that after implementation of the identified actions the cooling system could be operated in higher reliability and availability at 97,7% as required by the project.

RAMI analysis for ITER fuel cycle system [6] has identified several failure modes with high risks, majority of which was removed by implementing risk reducing means. However, some most critical risks remain, e.g. several critical components of tritium plant, which are not easily replaced or repaired.

Up to date the ITER project is probably the one which achieved the biggest advance in systematic use of reliability and risk analysis methods for fusion device.

2.2 Approach used for W7-X

The Wendelstein 7-X (W7-X) is an optimized stellarator experiment which shall demonstrate the possibility to use such a system as a nuclear fusion power plant. The

project is in the assembly and preparation for commissioning phase at the Max-Planck-Institut für Plasmaphysik (IPP) in Greifswald, Germany. The quality of plasma Wendelstein 7-X will start operation step by step in 2014, the first plasma is expected in 2015.

IPP has decided to use RAMI approach for the W7-X. As W7-X at that time was already in manufacturing and assembly state it was too late to make significant design changes. Therefore it was decided to perform reliability analysis based on modelling of already existing systems and then provide recommendations for improvement of system reliability and availability. This approach is different from one used for ITER, where the overall ITER availability goal is “distributed” among the systems and is defined for the systems and components (top-down approach). For the W7-X, contrary, was decided to use “bottom-up” approach when existing system availability is estimated and improved. Ideally, full-scope analysis would enable to obtain overall W7-X availability as a summary of all systems availabilities. Having such complete model would allow to see how improvements of each system design, operation, maintenance, inspections etc. would improve both systems and overall W7-X availability.

For purpose of reliability/availability and risk analysis of W7-X probabilistic safety assessment methods were used.

3 Overview of Methods for Analysis

3.1 Main Methods for Assessment

To estimate risk a probabilistic safety assessment (PSA), which is typically used for nuclear power plants, can be applied for any hazardous systems, e.g. [7, 8]. PSA methodology integrates information about device design, operating practices, operating histories, component reliabilities, humans’ behaviour, thermal hydraulic device response, accident phenomena and potential environmental and health effects. PSA is widely used for estimation of safety and reliability of energy generating complex systems.

Fault tree analysis (FTA) together with event tree analysis (ETA) are two main tools in system analysis. Both methods include quantification part and visual representations of Boolean logic for accident sequences [9]. FTA is an analytical technique, whereby an undesired state of the system is specified (usually a state that is critical from a safety or reliability standpoint), and the system is then analyzed in the context of its environment and operation to find all realistic ways in which the undesired event (top event) can occur. The fault tree is a graphic model of the various parallel and sequential combinations of faults, caused by hardware failures, human errors, software errors, or any other pertinent events, that will result in the occurrence of the predefined undesired state. The FTA attempts to develop a deterministic description of the occurrence of an event, called the top event, in terms of the occurrence or non-occurrence of other (intermediate) events. Intermediate events are also described further until the lowest level of detail, the basic events, are reached.

A fault tree analysis may be qualitative, quantitative, or both, depending on the objectives of the analysis. Possible results from the analysis may, for example, be:

- A listing of the possible combinations of environmental factors, human errors (if included), normal operational events, and component failures that may result in a critical state of the system.
- The probability that the critical event will occur during a specified time interval.

As a result of the fault tree initial qualitative analysis minimal cut sets (MCS) are generated. Minimal cut set is a set of basic events which, if occurred, definitely lead to the top event. Minimal cut set is a cut set such that after removal of any basic events from it is no more a cut set. When the fault trees are structured, the MCS generations and quantification for quantitive analysis is made by PSA software.

3.2 Importance and Sensitivity

In order to better understand the influence of each component and each parameter on the total system reliability/unavailability and risk the importance and sensitivity analyses are performed. The importance measures are:
The Fussell-Vesely (FV) importance for a basic event is the ratio between the unavailability based only on all MCSs where the basic event *i* is included and the nominal top event unavailability is:

$$I^{FV}{}_i = \frac{Q_{TOP}(MCS_{including}\, i)}{Q_{TOP}}; \tag{2}$$

where $I^{FV}{}_i$ – FV importance;
Q_{TOP} - nominal top event unavailability; Q_{TOP}(MCS including *i*) - unavailability based only on MCSs where the basic event *i* is included.

The risk decrease factor (RDF) is calculated as:

$$I^{D}{}_i = \frac{Q_{TOP}}{Q_{TOP}(Q_i = 0)}; \tag{3}$$

where $I^{D}{}_i$ – RDF; $Q_{TOP}(Q_i=0)$ – top event unavailability where unavailability of the basic event *i* is set to zero (the basic event does not contribute to the top event unavailability).

The risk increase factor (RIF) is calculated as:

$$I^{I}{}_i = \frac{Q_{TOP}(Q_i = 1)}{Q_{TOP}}; \tag{4}$$

where $I^{I}{}_i$ – RIF; $Q_{TOP}(Q_i=1)$ – top event unavailability where unavailability of the basic event *i* is set to one (the basic event does contribute to the top event unavailability).

The fractional contribution (FC) is calculated as:

$$I^{F}{}_i = 1 - \frac{1}{I^{D}{}_i}; \tag{5}$$

The sensitivity *S* is calculated as:

$$S = \frac{Q_{TOP,U}}{Q_{TOP,L}}; \tag{6}$$

where $Q_{TOP,U}$ – top event results where unavailability of the basic event *i* is multiplied by a sensitivity factor (normally equal to 10); $Q_{TOP,L}$ – top event results where unavailability of the basic event *i* is divided by the sensitivity factor;

4 Case Study: Wendelstein 7-X divertor target cooling circuit

The divertor target cooling circuit is a part of the water cooling circuits for the W7-X. It provides cooling flow for the target modules during plasma operation and ensures water circulation during other operational modes. It also provides heating up of the divertor target modules up to 150° C (so-called baking mode) before starting operation campaign after outage for maintenance.
The cooling circuit consists of a primary part (ACK10 cooling circuit) and a secondary part (ECB10 water supply system). The circles are separated by two parallel heat exchangers. The secondary part cools the primary part during the experiment and holds its temperature constant.
The primary part includes a cooling circuit and a separate baking circuit with its own pump and provides water to 110 parallel target modules.
The pipes with diameters of 25-600, the valves and other components are stainless steel. Water for the primary part must deionised. The total water content of the primary part is about 87 m³. The content of a lockable target module is about 25l.

Simplified flow diagram of the DTCC is provided in the following Figure 1.

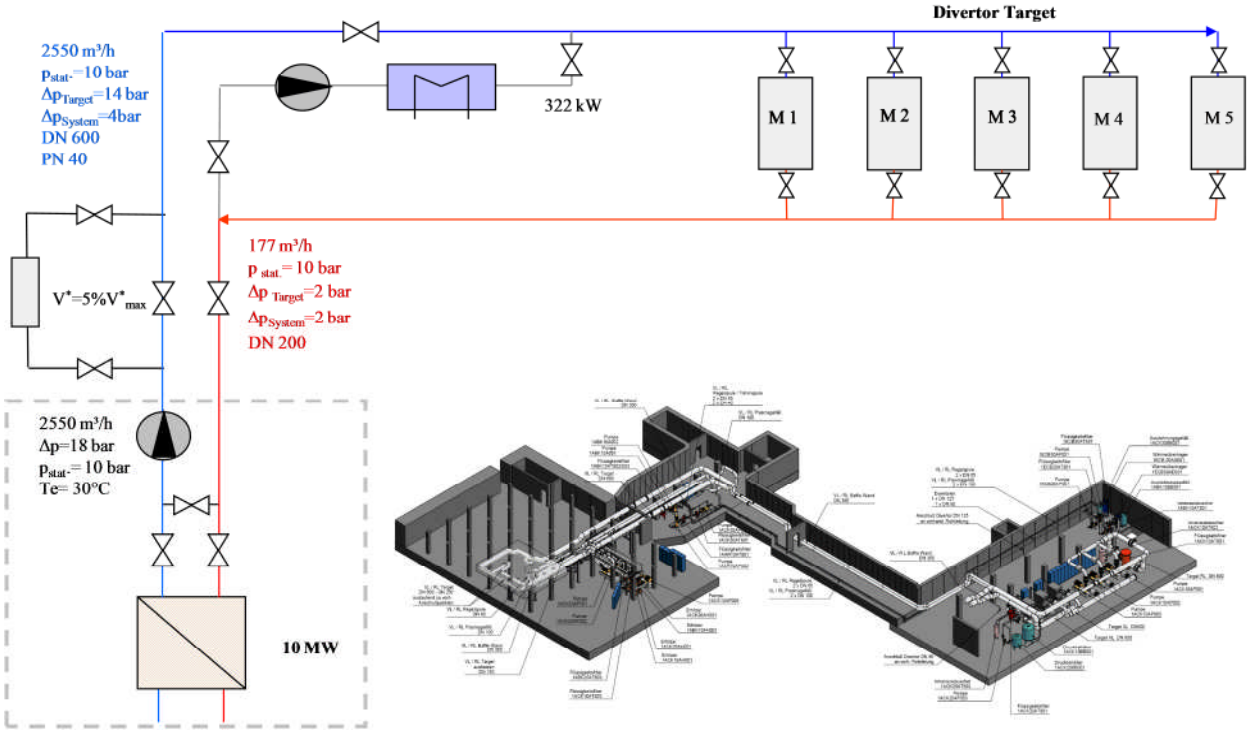


Figure 1. Simplified flow diagram and 3D schema of W7-X DTCC

Reliability analysis of the DTCC (ACK10) was performed using FTA and RiskSpectrum PSA software. The developed FTA model quantification includes minimal cutsets generation and uncertainty and sensitivity analysis. The calculation was performed for time period of 6526 hours i.e. total time of operation campaigns per one year. The MCS generations for analysis are made by PSA software.

Calculated total unavailability for ACK10 operation period in a year is 0.188. This means that the system will be unavailable for operation 18,8% of operation campaign. As a result of ACK10 model analysis 56 minimal cut sets were generated. The most important 7 MCS (which gives the highest unavailability) are presented in the following Table 1; the remaining MCS bring only 0.01% to the total unavailability.

No	Unavailability	% total	Event Name
1	9,57E-02	51	Pump AP002 fails to start
2	6,60E-02	35,1	Pneumatic valve KA510 Fails to open
3	2,11E-02	11,3	Pump AP003 fails to run
4	6,44E-03	3,43	Heat Exchanger AD002 fails
5	6,44E-03	3,43	Heat Exchanger AD001 fails
6	3,52E-03	1,87	Check valve KA507 Fails to open
7	4,50E-04	0,24	Pump AP002 fails to run

Table 1. Operation time for each ACK10 operation modes

The results show that due to low system redundancy the failure of a single component may lead to complete system unavailability. More than 50% influence to the system unavailability brings the failure of auxiliary (secondary) cooling pump AP002 which has a high quantity of cyclic loads. About 35% brings pneumatic valve KA510 located at the pressure line of the same pump. This valve is also a subject to high cyclic loads.
In order to better understand the influence of each component and each parameter on the total unavailability the importance and sensitivity analyses were performed.
The results of importance and sensitivity analyses for the most important 7 basic events are presented in the following Table 2.

No	Normal value	FV	FC	RDF	RIF	Sens	Sens high	Sens low
1	9,57E-02	5,10E-01	4,58E-01	1,84E+00	5,33E+00	8,71E+00	9,61E-01	1,10E-01
2	6,60E-02	3,51E-01	3,06E-01	1,44E+00	5,33E+00	5,17E+00	7,04E-01	1,36E-01
3	2,11E-02	1,13E-01	9,35E-02	1,10E+00	5,33E+00	2,01E+00	3,46E-01	1,72E-01
4	6,44E-03	3,43E-02	2,80E-02	1,03E+00	5,33E+00	1,28E+00	2,35E-01	1,83E-01
5	6,44E-03	3,43E-02	2,80E-02	1,03E+00	5,33E+00	1,28E+00	2,35E-01	1,83E-01
6	3,52E-03	1,87E-02	1,53E-02	1,02E+00	5,33E+00	1,15E+00	2,14E-01	1,85E-01
7	4,50E-04	2,40E-03	1,95E-03	1,00E+00	5,33E+00	1,02E+00	1,91E-01	1,87E-01

Table 2. Basic Events Importance and sensitivity analysis results

It is obvious that the most important basic events are the same as in the minimal cut sets. The interesting outcome is that sensitivity measures show how total unavailability would change if reliability of each component changes. For example, RDF shows that assuming “perfect” pumps with failure probability equal to 0 (1st basic event), this would decrease ACK10 unavailability 1,84 times and Sens low shows that increasing pump’s reliability 10 times, ACK10 unavailability would be 11% against current 18,4%.
The results of importance and sensitivity analyses for parameters are presented in the following Table 3.

No	ID	Type	Normal value	FC	RDF	RIF	Sens	Sens high	Sens low
1	ONE MONTH	Tr	7,20E+02	9,96E-01	2,37E+02	5,33E+00	3,65E+01	8,02E-01	2,20E-02
2	PUMP_STB	r	1,47E-04	4,58E-01	1,84E+00	5,33E+00	5,07E+00	5,64E-01	1,11E-01
3	PV_FTO	r	9,81E-05	3,06E-01	1,44E+00	5,33E+00	3,59E+00	4,90E-01	1,36E-01
4	PUMP_FTR	r	3,00E-05	9,35E-02	1,10E+00	5,33E+00	1,85E+00	3,18E-01	1,72E-01
5	HE_FAIL	r	9,00E-06	5,63E-02	1,06E+00	5,33E+00	1,54E+00	2,74E-01	1,78E-01
6	CV_FTO	r	4,90E-06	1,53E-02	1,02E+00	5,33E+00	1,15E+00	2,13E-01	1,85E-01
7	ONE_DAY	Tr	2,40E+01	3,44E-03	1,00E+00	5,33E+00	1,03E+00	1,94E-01	1,87E-01
8	MV_SPC	r	9,19E-07	2,10E-03	1,00E+00	5,33E+00	1,02E+00	1,91E-01	1,87E-01
9	PUMP_A_FTR	r	6,25E-07	1,95E-03	1,00E+00	5,33E+00	1,02E+00	1,91E-01	1,87E-01
10	PV_SPC	r	9,19E-07	1,15E-03	1,00E+00	5,33E+00	1,01E+00	1,90E-01	1,88E-01
11	PV_SPO	r	9,19E-07	1,91E-04	1,00E+00	5,33E+00	1,00E+00	1,88E-01	1,88E-01
12	MV_FTO	q	1,00E-04	6,60E-07	1,00E+00	1,01E+00	1,00E+00	1,88E-01	1,88E-01
13	FILTER_FAIL	r	2,00E-06	6,22E-07	1,00E+00	1,00E+00	1,00E+00	1,88E-01	1,88E-01

Table 3. Parameters importance and sensitivity analysis results

The final results show that the most important contributor to the system reliability is not equipment failure rates, but one month time period for hardware repair or replacement which was assumed (parameter ONE_MONTH). Sens low shows that decreasing this time 10 times would result in ACK10 unavailability only 2,2%. The next important parameter is pumps standby failure rate (parameter PUMP_STBY) which is used for pump AP002 and which improvement 10 times would change ACK10 unavailability from 18,8% to 11.1%. Other results for considered parameters can be interpreted in the same way.

5. Summary and Conclusions

A reliability and risk analysis of Divertor Target Cooling Circuit ACK10 was performed applying PSA related methods. The analysis included data collection, development of fault tree model, failure modes and effects analysis, estimation of reliability parameters and unavailability calculations.

The most important results and conclusions are:

1. Unavailability of the ACK10 is 18,8% of the operational campaign, i.e. about 1,5 month of 8 month operation in a year the system would be unavailable thus causing unavailability to use W7-X for experiments.
2. The main impact to unavailability is an operational regime of the cooling pumps where one pump is always running to provide cooling during all operational modes and the second one is started only to provide additional cooling for plasma experiments. This cause high cyclic load and corresponding high failure probability to the secondary pump (unavailability 95.7% which is almost certainly once per year) and it's regulating valve (unavailability 66%, i.e. twice in three years). Unavailabilities of these components bring correspondingly 51% and 35% to the total unavailability.
3. The another major reason for unavailability is long repair time which is assumed one month accounting for the time required to deliver and repair the equipment at the manufacturer's site or procure the spares required for the repair. Limited redundancy of the equipment does not enable to continue operation while the components are being repaired.

Comparison of W7-X and ITER reliability and risk analysis shows that:
W7-X analysis uses FTA and FMECA which is similar to RBD-FMECA approach for ITER. W7-X has less possibilities to make design changes in comparison with ITER therefore it should concentrate on such risk prevention and mitigation measures which would require less intervention to already designed and installed systems, such as:

- Improvement of maintenance programme;
- Improvement of operating/maintenance procedures;
- Hardware or system configuration changes only for most safety important components.

Acknowledgments
This paper was prepared on the basis of work, which was carried out within the framework of the European Fusion Development Agreement and supported by the European Communities. It was also partly supported by the Research Council of Lithuania. And last but not least the authors wish to acknowledge the large support and valuable assistance provided by of Dirk Naujoks from the Max-Planck-Institut für Plasmaphysik, Greifswald, Germany.

References

1. Todinov M.T., Risk-Based Reliability Analysis and Generic Principles for Risk Reduction, *Elsevier Science & Technology Books* (2006).
2. Ladra D., Sanguinetti G. and Stube E., "Fusion power plant availability study", *Fusion Engineering and Design,* vol. 58–59, pp. 1117-1121, (2001).
3. Van Houtte D., Okayama K. and Sagot F., "ITER operational availability and fluence objectives," *Fusion engineering and Design,* vol. 86, pp. 680-683, (2011).
4. Pamela J. et. al., "Efficiency and availability driven R&D issues for DEMO," *Fusion Engineering and Design,* vol. 84, no. (2–6), pp. 194-204, (2009).
5. Van Houtte D., Okayama K. and Sagot F., "RAMI approach for ITER," *Fusion Engineering and Design,* vol. 85, pp. 1220-1224, (2010).
6. Okayama K., van Houtte D., Sagot F. and Maruyama S., "RAMI analysis for ITER fuel cycle system,"*Fusion Engineering and Design,* vol. 86, pp. 598-601, (2011).
7. Hu L. and Wu Y., "Probabilistic safety assessment of the dual-cooled waste transmutation blanket for the FDS-I," Fusion Engineering and Design, vol. 81, no. 8-14, pp. 1403-1407, (2006).
8. Cambi G., Cavallone G., Costa M. and Ciattaglia S., "Summary of NET Plant Probabilistic Safety Approach and Results by Means of ENEA Fusion Plant Safety Assessment (EFPSA)," Journal of Fusion Energy, vol. 12, no. 1/2, pp. 127-131, (1993).
9. Caporal R. and Pinna T., "Multiple Failure Accident Sequences for SEAFP Reactor," Journal of Fusion Energy, vol. 16, no. 1/2, pp. 45-53, (1997).

Aleatory Uncertainty in Power System Reliability Index Assessment

R. Billinton[1] W. Wangdee[2]
1 University of Saskatchewan, Canada
2 BC Hydro, British Columbia, Canada

Abstract

The primary conventional reliability indices used in power system planning are mathematical expectations associated with annual system outage times or outage consequences. These are important indices that have been used for decades. The inherent variability in the annual loss of load or energy due to aleatory uncertainty is not generally appreciated or understood. This paper illustrates the aleatory uncertainty associated with power system reliability evaluation in the generating capacity domain. Numerical results and probability distributions obtained by applying sequential Monte Carlo simulation to a small practical test system are used to illustrate the aleatory uncertainty associated with basic generating capacity reliability indices. The increasing trend to incorporating intermittent generating capacity in the form of wind power in electric power systems introduces a new dimension in generating capacity planning. Highly variable wind capacity behaves quite differently from conventional generating capacity and influences the aleatory uncertainty associated with the predicted reliability indices. This influence is illustrated in this paper using the system well-being concept which incorporates both system security and adequacy considerations in the reliability analysis.

1. Introduction

There is a wide range of available indices to assess the reliability of electric power systems. These indices are used as criteria in planning generating capacity and transmission and distribution system additions and reinforcements. Similar indices are used in past performance assessments of existing systems. There are two fundamentally different forms of uncertainty that exist when predicting the reliability indices associated with future power systems or scenarios. The most obvious form is epistemic uncertainty, which arises from a lack of information regarding the future models and parameters required in the analysis. Epistemic uncertainty is knowledge based and can be reduced by better information. The epistemic uncertainty associated with load forecasting is readily appreciated and conventionally incorporated in generating capacity planning studies.

Uncertainty associated with the basic reliability indices also arises due to the random behaviour of the components included in the analysis. This predicted performance variability is known as aleatory uncertainty. The primary conventional reliability indices used in power system planning are mathematical expectations associated with annual system outage times or outage consequences [1]. Indices such as the loss of load expectation (LOLE) or loss of

energy expectation (LOEE) have been used in generating capacity reliability assessment for over fifty years. The inherent variability in the annual loss of load or energy due to aleatory uncertainty is not generally appreciated or understood. This paper illustrates the aleatory uncertainty associated with power system reliability evaluation in the generating capacity domain. Numerical results and probability distributions obtained by applying sequential Monte Carlo simulation to a small practical test system are used to illustrate the aleatory uncertainty associated with basic generating capacity reliability indices. A sequential Monte Carlo simulation technique was utilized in this paper as this technique is ideally suited to the analysis of intermittent resources such as wind power and its framework can incorporate the chronological characteristics of wind.

The increasing trend to incorporating intermittent generating capacity in the form of wind power in electric power systems introduces a new dimension in generating capacity planning. Highly variable wind capacity behaves quite differently from conventional generating capacity and influences the aleatory uncertainty associated with the predicted reliability indices. This influence is illustrated in this paper using the system well-being concept which incorporates both system security and adequacy considerations in the reliability analysis.

2. Study System

The study system used in this paper is the Roy Billinton Test System (RBTS) [2] developed at the University of Saskatchewan. The total installed capacity is 240 MW. The capacities of the eleven generating units range from 5 to 40 MW. The load model was developed using a bottom-up approach in which customer sectors were modeled at the various load points and aggregated to produce the total system load [3]. The system peak load is 179.28 MW. The wind power added to the basic RBTS is considered to be located at three different wind sites and the wind turbine generator (WTG) unit capacities vary from 0.2 to 2.0 MW. Wind speed models have unique characteristics that are highly dependent on the site locations. Hourly wind speed data for a three year period obtained from the National Renewable Energy Laboratory [4] were used to develop auto-regressive moving average (ARMA) time series models [5] for each site. The basic wind parameters for each site are shown in Table 1, including the correlation between the three site wind profiles.

Wind Farm Sites (Name)	Site M1	Site M2	Site O1
Mean wind speed (m/s)	9.10	8.38	10.03
Standard deviation (m/s)	5.50	4.48	5.20
Geographical location	mountain	mountain	offshore
Correlation w.r.t. Site M1	1.00	0.85	0.05

Table 1. Basic Wind Speed Data

3. Study Scenarios

The studies conducted examine the impact of generation resource variability on the system reliability indices using the following five study scenarios.

Base Case: the basic RBTS

Case A1: the basic RBTS with the addition of a 10 MW conventional generating unit and the peak load increased to 1.0593 per unit (189.91 MW).
Case A2: the basic RBTS with the addition of a 20 MW conventional generating unit and the peak load increased to 1.1156 per unit (200.00 MW).
Case A3: the basic RBTS with the addition of three 9.0 MW wind farms located at the three wind sites shown in Table 1 and the peak load increased to 1.0593 per unit.
Case A4: the basic RBTS with the addition of three 21 MW wind farms located at the three wind sites shown in Table 1 and the peak load increased to 1.1156 per unit.

The Base Case scenario is used to create a datum for comparison purposes in the form of a generating capacity adequacy criterion expressed by the system loss of load probability (LOLP). Cases A1 and A2 show the effect of adding a conventional generating unit to the system. Cases A3 and A4 illustrate the effect of adding intermittent wind power generation. In order to maintain the reliability level at the reference point, the system load has to be increased when adding new generation so that the system reliability can be maintained at the same value as that in the original system at the 1.0 per unit load level. As shown in Table 2, the generating system reliability expressed by the loss of load probability (LOLP) is the same in the Base Case and in Cases A1 and A3. The peak load in Cases A1 and A3 is increased to 1.0593 per unit after adding the new generation in order to maintain the same LOLP as that of the Base Case. This increased peak load in Cases A1 and A3 with respect to the Base Case 1.0 per unit load level is designated as the effective load carrying capability [6] or the peak load carrying capability due to the additional generation. In a similar manner, the peak load was increased to 1.1156 per unit to provide a similar outcome in Cases A2 and A4. The LOLP is the same in the Base Case and in Cases A2 and A4. The addition of 27 MW of wind power capacity in the form of the three 9 MW wind farms at the noted sites provides the same effective load carrying capability [6] as one 10 MW conventional generating unit. In Case A4, a total of 63 MW of wind power capacity in the form of three 21 MW wind farms is required to replace the added 20 MW generating unit. In order to maintain the same reliability level, the system requires a larger amount of intermittent generation than the required amount of conventional generation. This is due to the fact that intermittent generation such as wind power is highly uncertain compared to conventional generation and the system requires more wind power generating capacity to cope with the greater uncertainty.

4. Generating System Adequacy Evaluation

A wide range of indices are used to assess the adequacy of electric power systems [1]. The following are the most commonly used indices in the assessment of generating system adequacy. These indices are expected values associated with the situation in which the load exceeds the available generating capacity.

LOLP = Loss of Load Probability (/year)
LOLE = Loss of Load Expectation (hours/year) = LOLPx8760
LOLF = Loss of Load Frequency (occurrences/year)
LOEE = Loss of Energy Expectation (MWh/year)

The numerical values of the above indices for the five scenarios are shown in Table 2.

Base Case: Original system without wind power generation
Case A1: Add a 10 MW conventional unit @ 1.0593 p.u. load level
Case A2: Add a 20 MW conventional unit @ 1.1156 p.u. load level
Case A3: Add 3×9 MW wind farms @ 1.0593 p.u. load level
Case A4: Add 3×21 MW wind farms @ 1.1156 p.u. load level

Index	Base	Case A1	Case A2	Case A3	Case A4
LOLP	0.00043	0.00043	0.00043	0.00043	0.00043
LOLE	3.78	3.78	3.78	3.78	3.78
LOLF	0.80	0.82	0.75	0.95	1.02
LOEE	35.87	35.98	36.94	36.13	39.61

Table 2 Generation adequacy indices for the five generation scenarios

As noted earlier, and shown in Table 2, each scenario has the same *LOLP* and *LOLE*. The *LOLF*, however, for Cases A3 and A4, in which wind power is added, are higher than in the no wind cases. This implies that a system containing intermittent capacity such as wind power is likely to encounter a loss of load situation more frequently than in a system containing conventional generation [7] such as the RBTS. The *LOLF* in Cases A1 and A2 are similar to that for the Base Case. Table 2 also indicates that the *LOLE* tends to rise with increased wind power penetration.

The annual indices shown in Table 2 are long run average or expected values. The aleatory uncertainty associated with the performance in a given year can be described by a probability distribution. Figure 1 shows the probability distributions for the five generation scenarios based on the annual duration of load loss. The variate in this figure is designated as the Loss of Load Time (hours/year) and the distribution mean values are the *LOLE* indices shown in Table 2. The Loss of Load Incidence (occurrences/year) and Unserved Energy

(MWh/year) distributions associated with the *LOLF* and *LOEE* are shown in Figures 2 and 3 respectively.

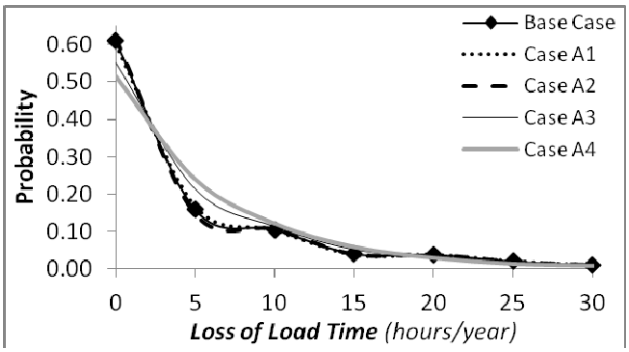


Figure 1: Probability distributions of the annual loss of load time for the five generation scenarios

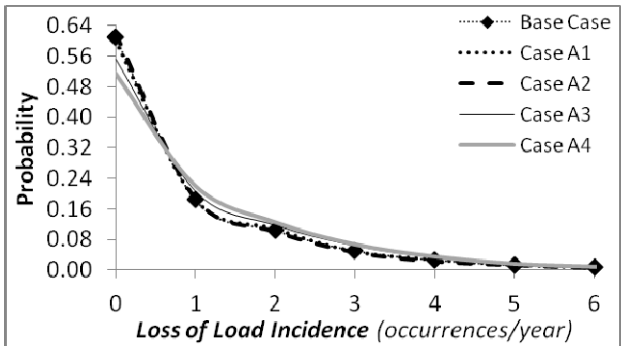


Figure 2: Probability distributions of the annual loss of load incidence for the five generation scenarios

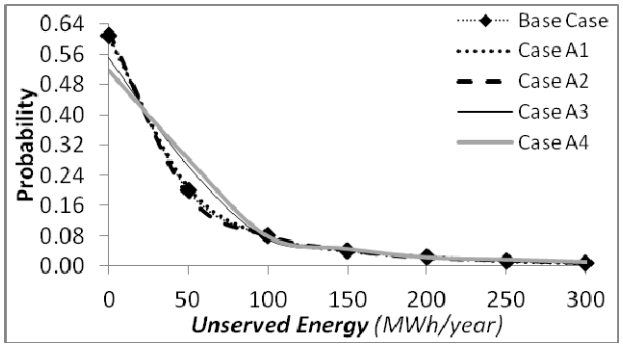


Figure 3: Probability distributions of the annual unserved energy for the five generation scenarios

The probability distributions are presented as approximate continuous distributions rather than histograms to facilitate a comparison of the scenario performances.

Figure 1 shows that while the *LOLE* is the same for all five cases, the uncertainty profiles are different. The probability distributions for the Loss of Load Time for the Base Case and Cases A1 and A2 in which a conventional unit is added are very similar but are different from those for Cases A3 and A4, which have wind power additions. It can be seen from Figure 1 that the probability of having no loss of load in a given year is lower for Cases A3 and A4. In other words, the wind integrated systems are more likely to encounter loss of load situations (Loss of Load Time >0) than the systems with no wind power additions

.

This phenomenon is further illustrated in Figures 2 and 3. The aleatory uncertainty associated with the generating system adequacy indices is not normally evaluated or discussed and the primary focus is placed on the expected values shown in Table 2.

5. Generating System Security Constrained Adequacy Evaluation

Power system reliability assessment can be generally divided into the fundamental designations of adequacy and security evaluation [1]. System adequacy relates to the existence of sufficient facilities within the system to satisfy the customer load demand. System security involves the ability of the system to respond to disturbances arising within the system. Table 2 illustrates the basic outcomes of a traditional generating capacity adequacy analysis in which the focus is on the loss of load condition. This traditional framework can be extended by incorporating a security constraint from a system operating point of view that considers how well the system is performing. System well-being analysis is a combined framework that incorporates a deterministic consideration in a probabilistic assessment to quantify the system conditions [8, 9].

The system well-being designated by the accepted deterministic criterion is categorized as being healthy (secure), marginal (insecure) and at risk (inadequate). The loss of the largest on-line generating unit is used in this study as the deterministic criterion. The system is in the healthy state when it has enough capacity reserve to meet the deterministic criterion. The system is not in difficulty in the marginal state but does not satisfy the deterministic criterion. The system is in the at risk state if the load exceeds the available capacity. The probability of the at-risk state is the traditional *LOLP*. The basic well-being system analysis model is shown in Figure 4.

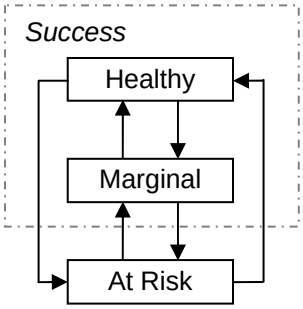


Figure 4: System well-being analysis framework

As previously noted, sequential Monte Carlo simulation was used in the system well-being analysis presented in this paper. The amount of generating capacity reserve required in the healthy state is determined by the capacity of the largest operating (online) unit at a particular point in time. This implies that the capacity of the largest operating generating unit may not be the same under different generating system states. Throughout the total period of study (i.e. a year), the generating capacity reserve is measured against the capacity of the largest operating unit at each particular hour to determine the health, margin and at-risk states. The degree of system well-being can be quantified in terms of the probabilities, frequencies and durations of the healthy, marginal and at-risk states. The probabilities and frequencies illustrated in this paper are defined as follows:

P_H = Healthy state probability (/year)
P_M = Marginal state probability (/year)
P_R = At-risk state probability (/year)
F_H = Healthy state frequency (occurrences/year)
F_M = Marginal state frequency (occurrences/year)
F_R = At-risk state frequency (occurrences/year)

System well-being analysis captures both adequacy and security concerns and provides increased insight on the overall reliability impacts of integrating intermittent wind power generation in conventional generating systems. Table 3 presents the system well-being indices for the five generation scenarios. The P_R value shown in Table 3 is the traditional *LOLP* shown in Table 2. The results in Table 3 indicate that while maintaining the system adequacy level at a P_R of 0.00043, the P_H of Cases A3 and A4 associated with wind power generation are lower than the Base Case P_H. This indicates that the system security level is not retained by adding only wind generation to maintain the specified adequacy level. Table 3 also indicates that the frequency indices (F_H, F_M and F_R) increase in Cases A3 and A4 compared to those in the Base Case, which shows that the systems states are more dynamic when adding wind power generation as there are more movements between the healthy and marginal states compared to the Base Case. In other words, the system containing wind generation can move more frequently to the marginal state where the system is still adequate but not

as secure as before. This conclusion is, however, not applicable in Cases A1 and A2 where a conventional unit is added. In these cases, the F_H, F_M and F_R are retained or improved. In addition, the P_H of Cases A1 and A2 are slightly higher than the Base Case P_H. This implies that the system security level can be retained (even further improved) when adding conventional generation to maintain the specified adequacy level.

Base Case: Original system without wind power generation *Case A1: Add a 10 MW conventional unit @ 1.0593 p.u. load level* *Case A2: Add a 20 MW conventional unit @ 1.1156 p.u. load level* *Case A3: Add 3×9 MW wind farms @ 1.0593 p.u. load level* *Case A4: Add 3×21 MW wind farms @ 1.1156 p.u. load level*					
Index	**Base**	**Case A1**	**Case A2**	**Case A3**	**Case A4**
P_H	0.98456	0.98462	0.98511	0.97848	0.98053
P_M	0.01501	0.01495	0.01446	0.02109	0.01904
P_R	0.00043	0.00043	0.00043	0.00043	0.00043
F_H	25.10	25.67	22.44	36.49	35.30
F_M	25.84	26.43	23.14	37.39	36.14
F_R	0.80	0.82	0.75	0.95	1.02

Table 3: System well-being indices for the five generation scenarios

The results shown in Table 3 are the average or expected values of the well-being indices. The system well-being index probability distributions, which provide a pictorial representation of the annual aleatory uncertainty of the indices, are illustrated in Figure 5. In order to differentiate the representation of the expected values of the system well-being indices in Table 3 from the representation of the system well-being index probability distributions, the symbols used to represent the annual variability (probability distribution) of the system well-being indices are as follows:

HST = Annual healthy state time (hours/year)
MST = Annual marginal state time (hours/year)
RST = Annual at-risk state time (hours/year)
f_H *= Annual healthy state frequency (occurrence/year)*
f_M *= Annual marginal state frequency (occurrence/year)*
f_R *= Annual at-risk state frequency (occurrence/year)*

The *HST*, *MST* and *RST* indices in Figure 5 were obtained by multiplying the relevant system state probability by 8760 in order to present the indices as state times in hours per year. The *RST* and f_R profiles presented in Figure 5 are the same as the profiles of Loss of Load Time (Figure 1) and Loss of Load Incidence (Figure 2) obtained in the traditional generation adequacy evaluation.

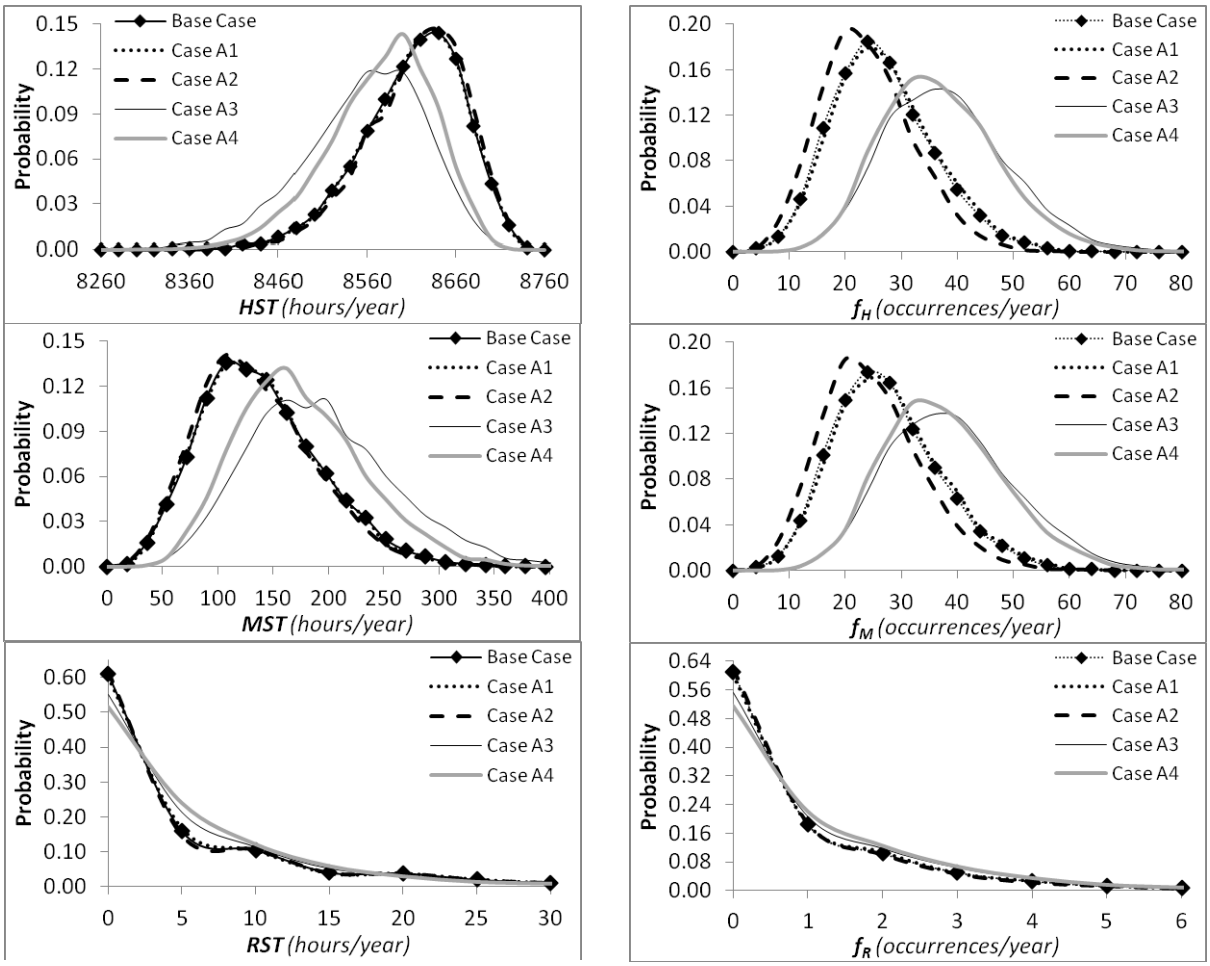


Figure 5: Probability distributions of the system well-being indices for the five generation scenarios

Figure 5 indicates that while the expected values of P_R for all the five cases are the same (P_R = 0.00043), the uncertainty distribution profiles of *HST* and *MST* for all the five cases are quite different. As previously noted, the system security level is not retained in a wind integrated power system, while maintaining the specified adequacy level. Figure 5 clearly shows that the *HST* distribution profiles of Cases A3 and A4 shift to the left (security level degradation) compared to those of the Base Case. Figure 5 also shows that the f_H distribution profiles for the systems containing wind power generation (Cases A3 and A4) shift to the right (departing from the secure state more frequently) with more dispersion and, therefore, greater uncertainty with lower predicted probabilities of occurrence compared to the Base Case. This conclusion is, however, not the case when adding a conventional generating unit (Cases A1 and A2) where the *HST* and f_H distribution profiles are quite similar to those of the Base Case indicating that the uncertainty level may remain relatively constant when adding conventional generation.

6. Conclusions

This paper illustrates how the traditional generating capacity indices of Loss of Load Expectation and Loss of energy Expectation can be complemented with a quantitative appreciation of the underlying aleatory uncertainty associated with the random variation in the annual generating capacity adequacy.

This appreciation is particularly important when considering the impact on the long term generating capacity adequacy of increased penetration of renewable energy sources such as wind power. The system well-being analysis illustrated in the paper is an important extension of the conventional approach to generating system adequacy assessment. System well-being index probability distribution analysis provides a visual representation of a multi-dimensional risk assessment approach that complements the single risk dimension provided by a expected value.

7. References

1. R. Billinton and R.N. Allan, “Reliability Evaluation of Power Systems”, 2nd Edition, New York, Plenum, 1996.
2. R. Billinton, et al, “A Reliability Test System for Educational Purposes – Basic Data”, IEEE Trans. Power Syst., vol. 4, no. 3, pp. 1238-1244, Aug. 1989.
3. A Sankarakrishnan and R. Billinton, “Sequential Monte Carlo Simulation for Composite Power System Reliability Analysis with Time Varying Loads”, IEEE Trans. Power Syst., vol. 10, no. 3, pp. 1540-1545, Aug. 1995.
4. Wind Integration Datasets – Western Wind Dataset. National Renewable Energy Laboratory (NREL) [online], Available: http://www.nrel.gov/wind/integrationdatasets/western/methodology.html.
5. R. Billinton, H. Chen and R. Ghajar, “Time-series Models for Reliability Evaluation of Power Systems Including Wind Energy”, Microelectronics and Reliability, vol. 36, no. 9, pp. 1253-1261, Sept. 1996.
6. L.L. Garver, “Effective Load Carrying Capability of Generating Units”, IEEE Trans. Power App. Syst., vol. PAS-85, no. 8, pp. 910-919, Aug. 1966.
7. W. Wangdee and R. Billinton, “Considering Load Carrying Capability and Wind Speed Correlation of WECS in Generation Adequacy Assessment”, IEEE Trans. Energy Convers., vol. 21, no. 3, pp. 734-741, Sept. 2006.
8. R. Billinton and M. Fotuhi-Firuzabad, “Basic Framework for Generating System Operating Health Analysis”, IEEE Trans. Power Syst., vol. 9, no. 3, pp. 1610-1617, Aug. 1994.
9. W. Wangdee and R. Billinton, “Probing the Intermittent Energy Resource Contributions from Generating Adequacy and Security Perspectives”, IEEE Trans. Power Syst., vol. 27, no. 4, pp. 2306-2313, Nov. 2012.

Choosing the reliability approach - A guideline for selecting the appropriate reliability method in the design process

Cristina Johansson[1,2], Per Persson[2], Michael Derelöv[1], Johan Ölvander[1]
Department of Management and Engineering, Linköping University, Sweden
SAAB Aeronautics, Bröderna Uggla Gatan, Linköping, Sweden

Abstract
The main objective of a reliability study should always be to provide information as a basis for decisions e.g. concept choice, design requirements, investment, choice of suppliers, design changes or guaranty claim. The choice of reliability method depends on the time allocated for the reliability study, the design stage, the problem at hand and the competence and resources available.

During a reliability study the engineer focuses on providing a graphical means of evaluating the relationships between different parts of the system, gather or assess the reliability data for the components and interpret the results of the analyses. Even though the commercial software tools available claim to provide answers to most reliability questions, the choice of which method that is best suited is not an easy task. Often several methods can be applied and none of them will fit the purpose perfectly.

This paper presents a guideline for choosing the best suited reliability method in early design phases, from two aspects: objective and system characteristics. The methods studied are the most common methods available in commercial software tools: Reliability Block Diagram (RBD), Fault Tree (FT), Event Tree (ET), Markov Analysis (MA) and Stochastic Petri Network (SPN). The guideline considers two aspects, the characteristics of the system studied, and the scope of the analysis. The applicability of each of the five chosen methods is assessed for all possible combinations of system characteristics and objective. A study has been done on Saab Aeronautics in order to evaluate the practical use of the analysed methods and how this guideline can improve the selection of appropriate reliability method in early design phases.

1. Background and Scope

The main objective of a reliability study should always be to provide information as a basis for decision making e.g. concept choice from a reliability point of view, design requirements such as redundancy, functional redundancy, protections, warnings, choice of suppliers, design changes in order to meet the safety and reliability requirements, guaranty claim, maintenance strategies and investments etc. Traditionally, reliability engineering focuses on critical hardware parts of the system and most of the reliability methods have been developed accordingly. The choice of method for reliability depends on the design schedule, the problem to solve and competence and resources allocated. Depending on the industry, several standards such as (IEEE, 1998) and (IEC 60300-3-1, 2003), or standards issued by organizations like International Standardization Organization (ISO) and European Commission for Space Standardization (ECSS), procedures and guidelines as for example (NASA, 1990) and (FAA, 2000) are available, in order to outline a standard practice for conducting reliability studies. Even though there are standards and handbooks available, the choice of the best fitting reliability method is still not an easy task. Often several methods can be applied and none of them will fit perfectly.

Several traditional reliability methods are available, and are incorporated in commercial software tools as well as “in house” tools. The commercial software tools developers claim to give you answers to most questions while the researchers try to solve complex reliability problems by using new mathematical methods or new methodologies. But from an engineering point of view, the study must give reasonable answers as quick as possible with a minimum effort and invested time.

While the research focus within the reliability field is on the mathematical modelling, during a reliability study, the engineer focuses on providing a graphical means of evaluating the relationships between different parts of the system. The confidence of the answers he gets, depend on the assumptions, quality of input data and the applicability of the reliability method used. The quality of input data often depends on the vendor and it is difficult for the reliability engineer to influence. The choice of methods can on the other hand increase the confidence of the answers.

The scope of this paper is to create a short guideline for choosing the best fitting reliability method, based on the combination of *system characteristics* and the *objective.*

2. Method and Analysis

There are many reliability methods and models (Rausand & Hoyland, 2004) developed in order to achieve more reliable and safe systems, and described in international standards and handbooks. According to (IEC 60300-3-1, 2003) one way of the methods to be classified is with regard to their main purpose: *methods for fault avoidance* (such as *Parts derating and selection*, *Stress-Strength Analysis* and *Part Count*), *methods for estimation of measures for basic events* (such as *failure rate prediction, human reliability analysis HRA, statistical reliability methods* or *software reliability engineering*) and *methods for architectural analysis and dependability assessment (allocation)*. The last category includes Failure Mode and Effect (and Criticality) Analysis (U.S. Department of Defence, 1980) or (IEC 60812, u.d.), Event Tree Analysis (IEC 62502, 2010), Fault Tree Analysis (IEC 61025, u.d.) or (Stamatelatos, et al., 2002), Zonal Analysis and Common Mode Fault (Federal Aviation Administration, 2000), Preliminary Hazard Analysis and Fault Hazard Analysis (MIL-STD-882D, 2000), Markov Analysis (IEC 61165, u.d.) and (International Electro-technical Commision, 2003), Petri Net Analysis (IEC 62551, 2012) or (ISO/IEC 15909, u.d.), Reliability Block Diagram (IEC 61078, u.d.), Common Cause Failure (Federal Aviation Administration, 2000) and (International Electro-technical Commision, 2003), and the list can continue.

The reliability methods considered in this paper are classical *methods for architectural analysis and dependability assessment* and the most common implemented in commercial software tools, such as: Reliability Block Diagram (RBD)- adopted for example to evaluate the reliability of three designs at both the functional and component level (O'Halloran, 2011), Fault Tree (FT)- one of the most popular method, used in many different applications, for example recently used to develop and analyse safety/security requirements for a gateway software (Kornecki & Liu, 2013), Event Tree (ET)- used often in early design phase in many applications, as for example to highlight common hazards arising from hydrogen storage and distribution systems, as well as to reveal potential accidents that hydrogen may yield under certain conditions (Rigas & Sklavounos, 2005), Markov Analysis (MA) and Stochastic Petri Network (SPN)- used for example to calculate the availability of safety critical on-demand systems (Kleyner & Volovoi, 2010) . The chosen methods can be used in conceptual design as well as all other design phases of a product development.

The variables taken into consideration in this paper in order to determine the choice of method (RBD, FT, ET, MA or SPN) are the *system characteristics* and the *goal of analysis*. These variables are chosen by the author from every day engineering practice, with regard to the impact on a reliability study.

The *system characteristics* are general in order for the method to be applicable to technical systems from many different fields such as the automotive and the aircraft industry (military and commercial). The proposed guideline considers six characteristics where each characteristic could have two mutually exclusive properties.

- The system behaviour: static or dynamic. Direct, explicit relationships among components (data path, workflow, feedback, etc.) creates a static behaviour, while load-sharing, standby redundancy, interferences, dependencies, on-demand, cascade, and/or common cause failures, human factor, fault-coverage, growth, phased-mission systems, time dependent sequences or several states systems and components qualifies for dynamic behaviour.
- Type of system: prototype or serial. A prototype system is unique and used for gathering information for future use while in the case of a serial system there are several identical individuals. The reliability models incorporate predictions based on parts-count failure rates taken from historical data. If the system analysed is unique, there is very little or no failure data information that can be used. Performing a reliability study on a prototype system is one of the challenges within the reliability field and therefore always important to specify the type of the system analysed.
- The system parts type: mostly mechanical/electromechanical or electronic parts. The electronic parts are considered well defined by exponential distribution (no aging) while the electromechanical and mechanical parts may have a different distribution (aging). Hence systems of a more mechanical nature (valves, pumps, rotors, generators, etc.) will show different behaviour (non-constant failure rate) from the electronic parts (constant failure rate) such as sensors, protection devices, inductors, capacitors, etc.
- Repairable or un-repairable system: Repairable systems receive maintenance actions to renew or restore the failed components when the system fails. These actions have to be taken into consideration when assessing the system behaviour. When the system fails during operation and the components that fail are not restored, the system is considered un-repairable.
- Safety or non-safety critical system: A safety-critical system is a system whose failure or malfunction may result in severe damages or injuries of persons, environment or equipment. Those systems will require not only a classical reliability study but sometimes extended risk analysis with event and consequence analyse. The analysis of such a system follows standards and handbooks such as for example (Federal Aviation Administration, 2000), (National Aeronautics and Space Administration Jet Propulsion Laboratory, 1990), (MIL-STD-882D, 2000), etc.

The different reliability questions that could arise during product development are addressed as *objective.* By contrast with the *system characteristics* (mutual exclusive choices), the *objective* can have several answer in the same time. In the proposed method the different questions have been grouped into the following six areas.

- System Reliability/ Unreliability: Usually calculated as the probability over the systems life time, the system reliability is a quality question and therefore has to be answered for any type of product from a large range of industries (automotive, aeronautics, space, manufacturing,

etc.). In early design phases can assure a more effective requirement selection, cost- performance trade off and a base for decisions regarding redundancies and maintenance, while in later design phases will help in guarantee issues. The system can have several phases/states and all of them should be accounted when analysing system reliability.

- States Probabilities: A system can have several degraded states (graceful degradation) and can be of interest to calculate depending of the customer, mission, etc. The reliability definition can be different for customers as well as mission performed and therefore a system with the same states (and state probability of occurrence) can have different output for reliability. These questions are important to answer for products within the manufacturing industry, automotive, aeronautics, etc at least in early design phases.
- Failure Scenarios/ Probability of an unwanted event: This question applies when the unwanted event is identified (usually with other analysis techniques such as Preliminary Hazard Analysis, Functional Hazard Assessment (Federal Aviation Administration, 2000), (International Electro-technical Commision, 2003), (MIL-STD-882D, 2000), (Rausand & Hoyland, 2004), etc.) and the calculation of the probability of occurrence is wanted. In those cases analysis are done in order to break down the faults causing the occurrence of the unwanted event until the root causes of these faults are identified. The failure scenarios analysis are performed from early design phases to detailed design in order to eliminate, avoid or mitigate failures and mandatory for system safety critical systems.
- Failure Scenarios/ Consequences for given events: This question applies when the unwanted event is identified (usually with other analysis techniques such as Preliminary Hazard Analysis, Functional Hazard Assessment (Federal Aviation Administration, 2000), (International Electro-technical Commision, 2003), (MIL-STD-882D, 2000), (Rausand & Hoyland, 2004), etc.) and probabilities of occurrence for consequences are wanted. In those cases failure scenarios causing certain outcomes of the given unwanted event are followed and probabilities of occurrence can be calculated. These failure scenarios analyses are performed from early design phases to detailed design in order to justify the fulfilment of safety requirements, test the efficiency of the mitigations, barriers, etc.
- Mission Reliability/ Unreliability: The same product (for example an airplane, a car, an industrial robot, etc) can require different functions depending of the mission. Usually calculated as the probability over the mission time is important for a mission planning (in any field this concept is used such as aeronautics, space, automotive, etc). Performed in early design phases can assure a more effective equipment selection, cost- performance trade off and a base for decisions regarding redundancies and maintenance, while in later design phases will help in guarantee issues.
- System Behaviour or Qualitative Analysis: In early design phases, when very little or no failure data is available, the reliability (and safety) study is qualitative. Reliability methods can be applied in order to

gather information about the system behaviour or to break down the safety (and reliability) requirements. System weaknesses (like single failures causing occurrence of a hazard or certain cut sets) can be discovered from such analysis.

Each of the five classical methods chosen in this guideline (RBD, FT, MA, ET, SPN) are analysed in order to determine how well the method can be used with regard to the different *objective* and *system characteristics* listed above. This analysis is presented in Table 1 and Table 2.

The applicability of each method is graded from one to three points, where:

*** means that the method fits well,
** means that the method fits well, with some exceptions and
* means that the method does not fit very well but it is possible to apply, or when using as a qualitative method it fits well, but not when used as quantitative method.

Where the method is not recommended for certain *system characteristic* (Table 1), no point is accounted in the respective table. This will exclude the respective method from analysed scenarios.

For example, if the system analysed has a static behaviour, the application of Stochastic Petri Network will be a waste of time but other methods such as Reliability Block Diagram are easier to apply and fit better. In this case no scenario with SPN for a system with static behaviour will be analysed. However, if the system analysed has a dynamic or dependent behaviour, the RBD is not able to model the relationship between components (see Table 1) and this scenario is excluded from this analyse.

If the analysed system is a prototype system (see Table 1) none of the methods will fit very well. The reason for this is that historical data and field experience are missing, leading to uncertainty in the result of the analysis.

In order to decide what kind of system we are dealing with, we have to go through all the *system characteristics* from A to E (Table 1) as follows:

A. Have the system static or dynamic behaviour?
B. Is the analysed system a prototype or a serial system?
C. Is the system composed mostly by electromechanical/mechanical or electronic parts?
D. Are we dealing with a repairable or non-repairable system?
E. Is the system safety or non-safety critical?

In order to analyze all the scenario combinations for system characteristics, a tree structure is used, see Figure 1. A method is qualified to use if it is qualified for every single characteristic of the system. For example, in scenario number 2 in Figure 1, RBD is not qualified because it is not recommended for systems with dynamic behaviour, while FT is not qualified because it is not recommended for non-safety critical systems.

Characteristic (mutual exclusive) [1]		RBD	ET	FT	SPN	MA
A	Static behaviour	***	**	***		*
	Dynamic dependent behaviour		**	**	***	***
B	Prototype System	*	*	*	*	*
	Serial System	***	***	***	**	***
C	Mostly electromechanical/ mechanical parts	***	**	***	**	**
	Mostly electronic parts	*	**	**	***	***
D	Repairable system		*	*	***	***
	Un-repairable system	***	**	***	*	***
E	Safety Critical	*	***	***	***	***
	Non-safety Critical	***	*		***	***

Table 1: Choice of method considering different *System Characteristics*

A measure to grade the methods applicability for a certain system is defined by the cumulated points for all system characteristics for each method. This measure will have a minimum value of 5 points (considered poor fitting) and a maximum value of 15 points (excellent fitting). The following intervals are used to determine the quality of fit:

- 5 to 7 for poor fitting
- 8 to 12 for good fitting
- 13 to 15 for very good fitting.

In the Table 2 the choice of method is performed following the *objective* of the analysis. Where the method is not recommended for certain system characteristic (Table 2), no point is accounted in the respective table. This will exclude the respective method from analyzed scenarios, in the same way as in Table 1.

Objective	RBD	ET	FT	SPN	MA
System Reliability/ Unreliability	***				***
States Probabilities	*		*		***
Failure Scenarios/ Probability of an unwanted event			***		**
Failure Scenarios/ Consequences for given events		***			
Mission Reliability/ Unreliability	***				***
System behaviour- Qualitative analysis		***	***	***	***

Table 2: Choice of method considering different *Objective*

The *objective* can include several questions included in the question categories presented in the Table 2. The same reasoning about the points used in Table 1 is used as well in Table 2.

[1] The System is defined by Characteristics A to E, every each of them with two mutual exclusive possibilities

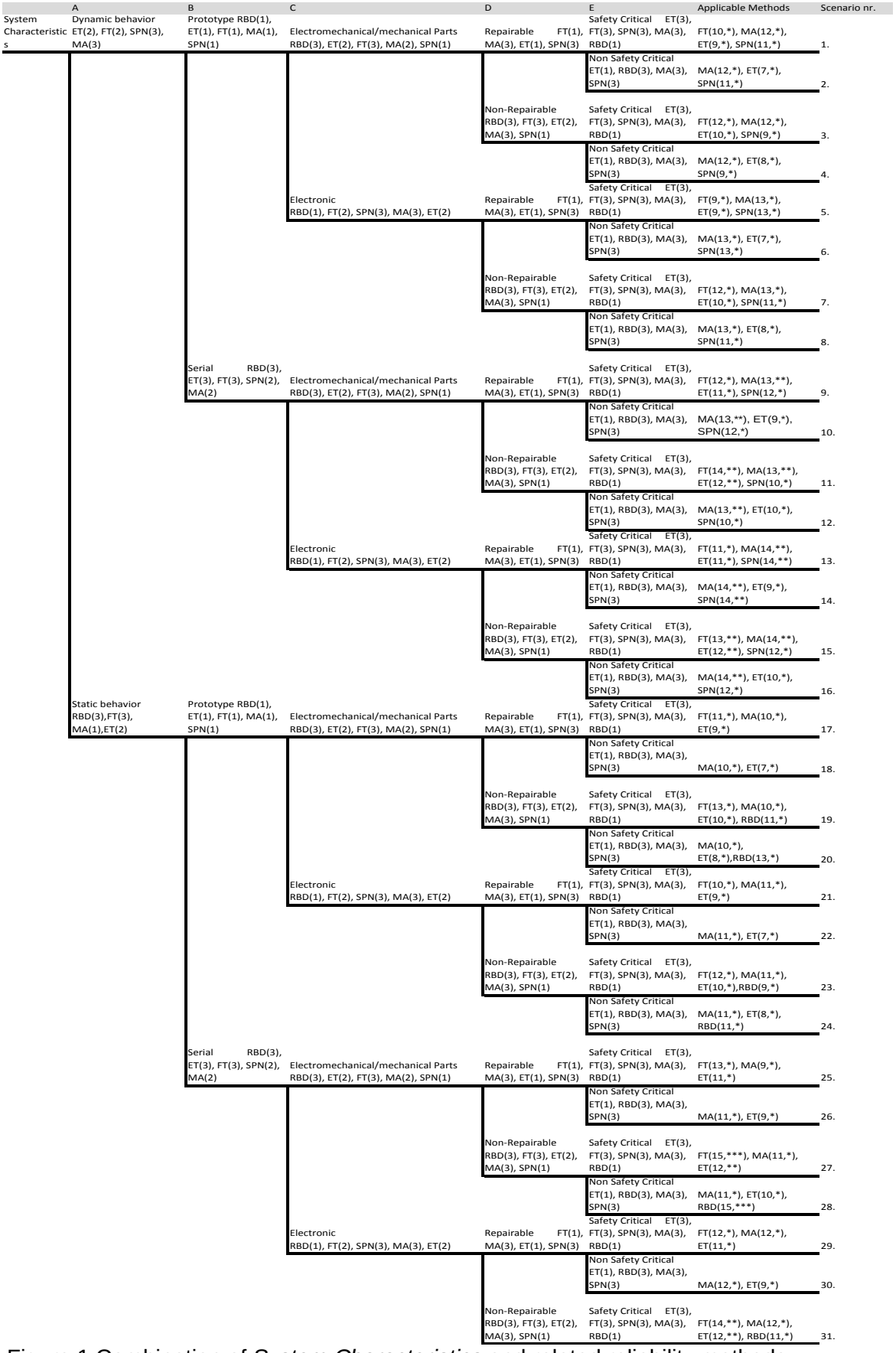


Figure 1 Combination of *System Characteristics* and related reliability methods

For example, if the *objective* is a question about system reliability (such as what is the reliability of the General Electronic Computer Unit), only two paths are available to the study, corresponding to RBD and MA methods. However, we have to ask all six questions listed in the Table 2 with answers of yes or no.

All the possible system scenarios derived from the matrix presented in Table 1 and analyzed in Figure 1, are combined with the objectives of the reliability/ system safety analysis from Table 2. Table 3 presents the fit of each reliability method for a certain objective and a system with certain characteristics. The engineer has to choose the scenario (1 to 32) describing the system to be analyzed and the objective of the analysis. One or more methods are suggested for use.

For example, if the engineer wants to know the System reliability or unreliability, for a safety critical, repairable, prototype system with dynamic behaviour, composed mostly of electromechanical/mechanical parts (row 1- first scenario in the Table 3), the recommended method is Markov Analysis which has a good fitting.

Scenario no. according to the Figure 1	**System Reliability / Unreliability**	**State Probabilities**	**Failure Scenarios / Probability of an unwanted event**	**Failure Scenarios / Consequences for given events**	**Mission Reliability/ Unreliability**	**System behaviour Qualitative analysis (barrier efficacy, sequence dependent failure scenario, etc)**
1	MA(*) 12	MA(*) 12	FT(*) 10, MA(*) 12	ET(*) 9	MA(*) 12	MA(*) 12, FT(*) 10, ET(*) 9, SPN(*) 11
2	MA(*) 12	MA(*) 12	MA(*) 12	ET(*) 7	MA(*) 12	MA(*) 12, ET(*) 7, SPN(*) 11
3	MA(*) 12	MA(*) 12	FT(*) 12, MA(*) 12	ET(*) 10	MA(*) 12	MA(*) 12, FT(*) 12, ET(*) 10, SPN(*) 9
4	MA(*) 12	MA(*) 12	MA(*) 12	ET(*) 8	MA(*) 12	MA(*) 12, ET(*) 8, SPN(*) 9
5	MA(*) 13	MA(*) 13	MA(*) 13, FT(*) 9	ET(*) 9	MA(*) 13	MA(*) 13, ET(*) 9, FT(*) 9, SPN(*) 13
6	MA(*) 13	MA(*) 13	MA(*) 13	ET(*) 7	MA(*) 13	MA(*) 13, ET(*) 7, SPN(*) 13
7	MA(*) 13	MA(*) 13	MA(*) 13, FT(*) 12	ET(*) 10	MA(*) 13	MA(*) 13, ET(*) 10, FT(*) 12, SPN(*) 11
8	MA(*) 13	MA(*) 13	MA(*) 13	ET(*) 8	MA(*) 13	MA(*) 13, ET(*) 8, SPN(*) 11
9	MA(**) 13	MA(**) 13	MA(**) 13, FT(*) 12	ET(*) 11	MA(**) 13	MA(**) 13, ET(*) 11, FT(*) 12, SPN(**) 12
10	MA(**) 13	MA(**) 13	MA(**) 13	ET(*) 9	MA(**) 13	MA(**) 13, ET(*) 9, SPN(**) 12

11	MA(**) 13	MA(**) 13	MA(**) 13, FT(**) 14	ET(**) 12	MA(**) 13	MA(**) 13, ET(**) 12, FT(**) 14, SPN(*) 10
12	MA(**) 13	MA(**) 13	MA(**) 13	ET(*) 10	MA(**) 13	MA(**) 13, ET(*) 10, SPN(*) 10
13	MA(**) 14	MA(**) 14	MA(**) 14, FT(*) 11	ET(*) 11	MA(**) 14	MA(**) 14 ET(*) 11, FT(*) 11, SPN(**) 14
14	MA(**) 14	MA(**) 14	MA(**) 14	ET(*) 9	MA(**) 14	MA(**) 14, ET(*) 9, SPN(**) 14
15	MA(**) 14	MA(**) 14	MA(**) 14, FT(**) 13	ET(**) 12	MA(**) 14	MA(**) 14, ET(**) 12, FT(**) 13, SPN(*) 12
16	MA(**) 14	MA(**) 14	MA(**) 14	ET(*) 10	MA(**) 14	MA(**) 14, ET(*) 10, SPN(*) 12
17	MA(*) 10	MA(*) 10	MA(*) 10, FT(*) 11	ET(*) 9	MA(*) 10	MA(*) 10, ET(*) 9, FT(*) 11
18	MA(*) 10	MA(*) 10	MA(*) 10	ET(*) 7	MA(*) 10	MA(*) 10, ET(*) 7
19	MA(*) 10, RBD(*) 11	MA(*) 10, RBD(*) 11	MA(*) 10, FT(*) 13	ET(*) 10	MA(*) 10 RBD(*) 11	MA(*) 10, ET(*) 10, FT(*) 13
20	RBD(*) 13, MA(*) 10	RBD(*) 13, MA(*) 10	MA(*) 10	ET(*) 8	RBD(*) 13, MA(*) 10	MA(*) 10, ET(*) 8
21	MA(*) 11	MA(*) 11	MA(*) 11, FT(*) 10	ET(*) 9	MA(*) 11	MA(*) 11, ET(*) 9, FT(*) 10
22	MA(*) 11	MA(*) 11	MA(*) 11	ET(*) 7	MA(*) 11	MA(*) 11, ET(*) 7
23	MA(*) 11, RBD(*) 9	MA(*) 11 RBD(*) 9	MA(*) 11, FT(*) 12	ET(*) 10	MA(*) 11 RBD(*) 9	MA(*) 11, ET(*) 10, FT(*) 12
24	RBD(*) 11, MA(*) 11	RBD(*) 11, MA(*) 11	MA(*) 11	ET(*) 8	RBD(*) 11, MA(*) 11	MA(*) 11, ET(*) 8
25	MA(*) 9	MA(*) 9	MA(*) 9, FT(*) 13	ET(*) 11	MA(*) 9	MA(*) 9, ET(*) 11, FT(*) 13
26	MA(*) 11	MA(*) 11	MA(*) 11	ET(*) 9	MA(*) 11	MA(*) 11, ET(*) 9
27	MA(*) 11, RBD(*) 13	MA(*) 11, RBD(*) 13	MA(*) 11, FT(***) 15	ET(**) 12	MA(*) 11, RBD(*) 13	MA(*) 11, ET(**) 12, FT(***) 15
28	RBD(***) 15, MA(*) 11	RBD(***) 15, MA(*) 11	MA(*) 11	ET(*) 10	RBD(***) 15, MA(*) 11	MA(*) 11, ET(*) 10
29	MA(*) 11	MA(*) 12	MA(*) 12, FT(*) 12	ET(*) 11	MA(*) 12	MA(*) 12, ET(*) 11, FT(*) 12
30	MA(*) 12	MA(*) 12	MA(*) 12	ET(*) 9	MA(*) 12	MA(*) 12, ET(*) 9
31	MA(*) 12, RBD(*) 11	MA(*) 12, RBD(*) 11	MA(*) 12, FT(**) 14	ET(**) 12	MA(*) 12, RBD(*) 11	MA(*) 12, ET(**) 12,FT(**) 14
32	RBD(*) 13, MA(*) 12	RBD(*) 13, MA(*) 12	MA(*) 12	ET(*) 10	RBD(*) 13, MA(*) 12	MA(*) 12, ET(*) 10

Table 3: Choice of method considering both *Scope of analyses* and *System Characteristics*

If several methods are recommended for the same scope of the analysis, the analyst can chose between the methods depending on fitting points, experience or if one of the methods can give the answers for several objectives. If the aim of the reliability study is the system behaviour, several methods can be chosen.

3. Application

As an example, the following questions are relevant to answer for a reliability study for an Electrical Power Supply System of an aircraft:

1. What is the mission reliability for the Electrical Power Supply function? Several phases need to be considered such as taxing, take off, flight and landing.
2. What are the probabilities of failure (failure rate) of safety critical functions? For example Emergency Power Supply, Auxiliary Power Supply, etc.
3. What are the probabilities of an initiating event to result in certain consequences? For example loss of aircraft due to total loss of AC power.
4. What is the probability of electrical power supply failure for certain consumers? For example loss of power supply to General Electronic Control Unit, loss of power supply to cockpit displays, etc.

The characteristics of the system according to Table 1 are:

A-dynamic, dependent behaviour;
B-serial system;
C- mostly electromechanical/ mechanical parts;
D- non repairable during flight;
E- safety critical.

The *objectives* are according to Table 2:

1. Mission Reliability/ Unreliability
2. Failure Scenarios/ Probability of an unwanted event,
3. Failure Scenarios/ Consequences for given events,
4. States Probabilities.

In Table 3 the scenario for the Electrical Power System is presented on row 11. The recommended methods are:

- Markov Analysis for question 1 and 4,
- Fault Tree or Markov Analysis for question 2 and
- Event Tree for question 3.

When the recommended methods are Fault Tree and/or Markov Analysis, the possibility of using dynamic fault tree gates for modeling certain dynamic behaviour should be investigated. This depends on what level of detail that is relevant for the questions asked and which design phase that is considered. The majority of commercial reliability software will support such dynamic gates.

When the choice is between two methods with the same fitting points, the choice will depend on other factors such as for example if a quick answer is more important than the accuracy of the answer (typical for concept phase).

4. Conclusions

This paper presents a guideline for choosing the best suited reliability method in early design phases, from two aspects: *system characteristics and objective*.

The guideline is deliberately written as general as possible to be applicable to many fields. Questions from the daily engineering practise are summarized in the Table 1 and Table 2. A decision tree (Figure 1) combines all the *system characteristics* (Table 1) in a number of possible systems, with respective fitted method to analyse. Finally, in the Table 3 these scenarios are combined with the *objective* from the Table 2. In the Table 3 at list one reliability method will be suggested to use, depending on what question is asked for the system to analyse.

The aspects analysed here has been chosen to be as general as possible and tested on different systems in order to verify the applicability of the guideline. However, the engineer will sometimes be forced to consider other aspects than those analysed, such as the capability of used reliability tool, field experience, time and resources allocated, etc.

There are some drawbacks such as the limited amount of methods considered (only five methods), and considerations regarding the system knowledge. The software reliability is not considered and neither is the failure data source and relevance. In the future work, several methods will be considered as well as a possible connection to the failure data.

References

1. Institute of Electrical and Electronics Engineers, 1998. *IEEE Standard Reliability Program for the Development and Production of Electronic Systems and Equipment.* s.l.:s.n.
2. Federal Aviation Administration, 2000. *FAA System Safety Handbook.* s.l.:s.n.
3. IEC 60812, n.d. *Analysis Techniques for system reliability - Procedure for FMEA.* s.l.:s.n.
4. IEC 61025, n.d. *Fault Tree Analysis (FTA).* s.l.:s.n.
5. IEC 61078, n.d. *Analysis techniques for dependability - Reliability block diagrams and boolean methods.* s.l.:s.n.
6. IEC 61165, n.d. *Application of Markov Techniques.* s.l.:s.n.
7. IEC 62502, 2010. *Analysis techniques for dependability – Event tree analysis (ETA).* s.l.:s.n.
8. IEC 60300-3-1, 2003. *Application Guide- Analysis techniques for dependability- Guide on methodology.* s.l.:s.n.
9. ISO/IEC 15909, n.d. *High-level Petri Nets - Concepts, Definitions and Graphical Notation.* s.l.:s.n.

10. Johansson, C., Persson, P. & Ölvander, J., 2011. *On the Usage of Reliability Methods in Early Design Phases.* Helsinki, PSAM11 & ESREL 2012.
11. Kleyner, A. & Volovoi, V., 2010. Application of Petri nets to reliability prediction of occupant safety systems with partial detection and repair. *ELSEVIER*, June.
12. Kornecki, A. & Liu, M., 2013. Fault Tree Analysis for Safety/Security Verification in Aviation Software. *Electronics,* Volume 2, pp. 41-56.
13. Lough, K., Stone, R. & Tumer, I., 2005. *THE RISK IN EARLY DESIGN (RED) METHOD: LIKELIHOOD AND CONSEQUENCE FORMULATIONS.* Philadelphia, ASME 2005 International Design Engineering Technical Conferences.
14. MIL-STD-882D, 2000. *Department of Defense Standard Practice For System Safety.* s.l.:s.n.
15. National Aeronautics and Space Administration Jet Propulsion Laboratory, 1990. *JPLD-5703 Reliability Analysis Handbook.* s.l.:s.n.
16. O'Halloran, B., 2011. *EARLY DESIGN STAGE RELIABILITY ANALYSIS USING FUNCTION-FLOW FAILURE RATES.* s.l., ASME 2011 International Design Engineering Technical Conferences .
17. Rausand, M. & Hoyland, A., 2004. *System Reliability Theory, Models, Statistical Methods and Applications.* Second ed. s.l.:Wiley.
18. Rigas, F. & Sklavounos, S., 2005. Evaluation of hazards associated with hydrogen storage facilities. *International Journal of Hydrogen Energy,* 30(13-14), p. 1501–1510.
19. Stamatelatos, D. M. et al., 2002. *Fault Tree Handbook with Aerospace Applications.* s.l.:s.n.
20. U.S. Department of Defence, 1980. *MIL-STD-1629A Procedures for Performing a Failure Mode Effects and Criticality Analysis.* s.l.:s.n.
21. IEC 62551, 2012. Analysis Techniques for Dependability-Petri Net techniques. s.I s.n

Investigating Electronics Reliability in Business Jet Applications

Ian James

Aero Engine Controls, Birmingham, UK

Abstract

Aero Engine Controls has developed a family of electronic control systems that are used on a wide range of aircraft platforms; large civil airliners, military jets and helicopters, right through to regional and business jets. To maximise design pedigree, much of the control system design is very similar, while the environment associated with each of the platforms on which they operate, may differ considerably.

1. Introduction

Business jet owners utilise their aircraft in a very different way to airline operators and service data suggests that this difference has a large effect on reliability. Aero Engine Controls has used its extensive knowledge and datasets associated with large civil operations to compare and contrast with aspects of business jet operation, in an attempt to understand and maximize product reliability.

This paper describes an investigation to understand the significance of this difference and its impact on reliability.

1.1 Aircraft Operation

Business jet aircraft manufacturers operate at the very high end of the aerospace market where quality and reliability are key drivers. If an aircraft becomes unavailable for operation at the time it is required, the consequence of delaying or cancelling the flight may be enormous; it is this reliability that business jet owners invest in. Consequently, reliability of all component parts is paramount. A key metric for the suppliers of component parts to the business jet operator is the number of flight cancellations that have been caused by failure of their equipment; this metric is often termed *Missed Trips*.

As civil airline operators manage many hundreds of flights per day, often with a mixed fleet of aircraft, simple reliance on component reliability is not sufficient to maintain flight operations. In order to minimise disruption to flight scheduling caused by faulty parts, airline operators will utilize a maintenance organisation with responsibility to manage the smooth running of the daily schedule. This organisation, which may be part of the airline business or a third party, will diagnose and correct faulty components between flights to maintain on-time departures.

As most business jet users will operate only a single aircraft, the use of a maintenance organisation is not a credible option so a far more proactive approach to component reliability must be taken. To this end, business jet manufacturers will monitor equipment reliability very closely and will expect shortfalls in performance to be dealt with in a timely fashion. They form very close relationships with engine manufacturers and their suppliers and will engage in regular reviews to discuss reliability performance.

1.2 Reliability Management

Reviews between the aircraft, engine and equipment manufacturers are held on a regular basis to monitor reliability performance and to agree remedial action. Any failures that have occurred within the reporting period are discussed in depth and the status of any ongoing investigations will be reviewed. Those failures that have caused missed trips will be given special attention, particularly with respect to containment action.

In the reliability reviews that Aero Engine Controls have been involved, Pareto analysis [1] has typically been relied upon to indicate which failures have the most impact upon product reliability. The results of this analysis will then be used to ensure remedial action is prioritised accordingly.

2. Reliability Improvement Program

It has long been recognised within Aero Engine Controls that the reliability of Electronic Control Units [ECUs] used on large aircraft appears superior to those installed on a business jet. ECUs used on large aircraft consistently demonstrate increased Mean Time Between Failure [MTBF] when compared with the business jet equivalent.

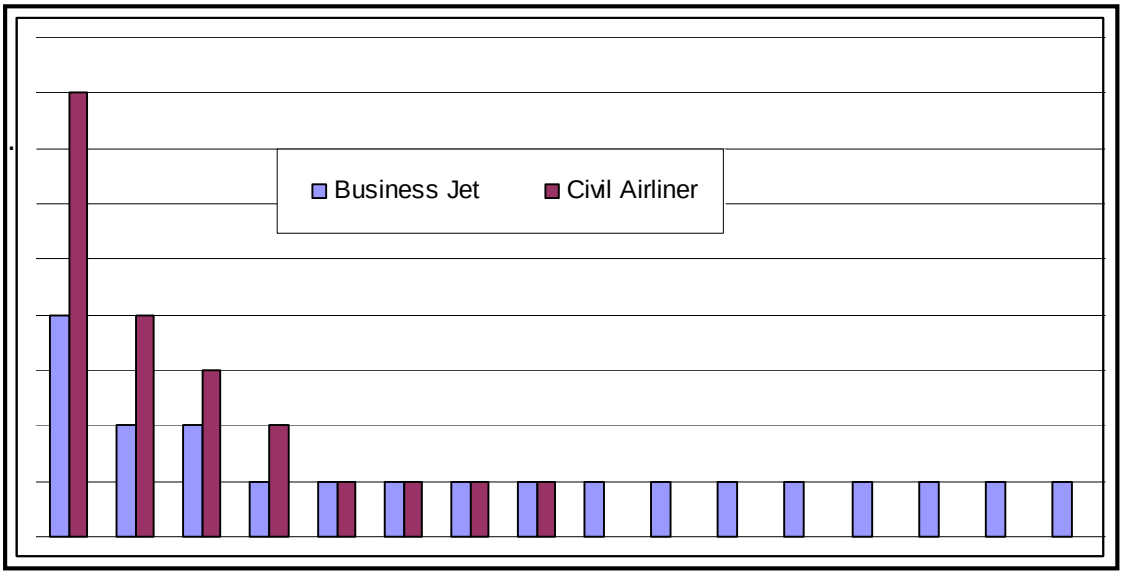


Figure 1: Difference in Failure Pareto for Business Jet and Civil Airliner

Prior studies to uncover potential reasons for this disparity have indicated very different failure characteristics. The failure Pareto shown in Figure 1 illustrates that an ECU used in a business jet application exhibits a much larger tail than an ECU used in a civil airliner. This may suggest that business jet ECUs suffer from many more, independent, failures than ECUs installed on a larger aircraft, which tend to exhibit a higher proportion of systematic failures.

The reliability of large aircraft ECUs can be improved significantly by modification action which remedies systematic faults. To realise a similar improvement in reliability for the business jet ECU, many modifications would be required; each with a small effect on reliability. To compound this problem, service data indicates the extent of the long tail has remained constant over time, indicating that ongoing modification action is not an effective management strategy.

This insight was discussed at length, during a regular reliability performance review with the engine and business jet manufacturer. During the review it was agreed that an alternative approach to reliability improvement should be adopted.

2.1 Reliability Task Team

In response to concerns expressed by the business jet manufacturer, that current reliability improvement initiatives were seen to be ineffective, an integrated reliability improvement program was initiated to specifically examine apparent ‘one-off’ failures. A cross-functional team represented by engineering staff from Aero Engine Controls, the engine manufacturer and the aircraft manufacturer was challenged to uncover systematic issues that may link these failures together.

The team’s objective was to define an appropriate investigation strategy which combines the experience of individual team members with a data collection and analysis process which uncovers the underlying common cause for apparent, multiple, one-off events.

2.2 Investigation Strategy

The team based their investigation strategy on a Physics of Failure [PoF] approach [2]. Exploratory Data Analysis [EDA] and accelerated test methods [3] were used to uncover elements of interest for further investigation while the Stress and Strength relationship [1] was employed as the basis for understanding the fundamental root cause of each electronic component failure. The team agreed it was important for them to bridge the gap between analysis and practical application. They decided the most effective means for instilling confidence in their analyses was to model the operational effect of each component failure, using Life Data Analysis [4,5], to allow direct comparison with observed behaviour from service experience.

The team defined three sources of data as prerequisite to carry out their analyses; Operator and Service reject data were requested along with failure data recorded during the ECU manufacturing process. The investigation strategy is shown in its entirety in figure 2.

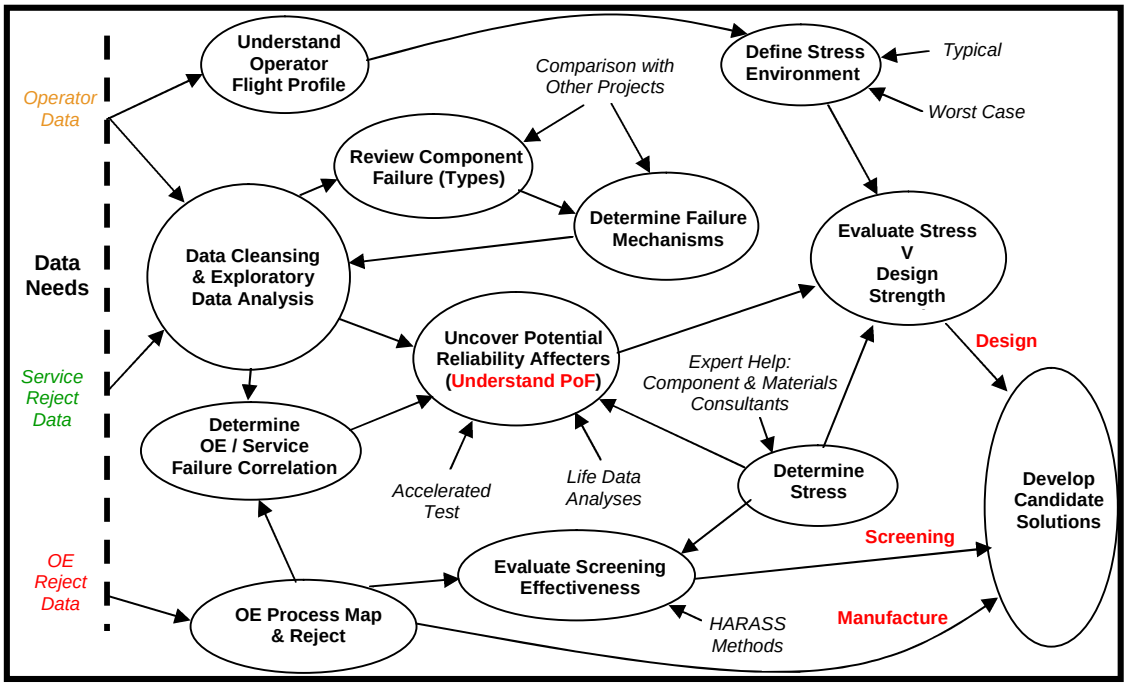


Figure 2. Investigation Strategy

3. Knowledge Management

From the outset of the investigation, the team recognised that if they were to employ an holistic approach successfully they must elicit, record and review data from multiple sources. It was agreed that the task team would visit the engine and aircraft manufacturer facilities to speak to domain experts and observe the engine and aircraft assembly process.

To initiate the knowledge capture process, Aero Engine Controls agreed to formally record and present their experience of large engine ECU systems, in terms of design, manufacture and usage envelope. The intention was to use this information as a prompt for other team members to verify, reject, modify or add to the data, based upon their collective experience. This evolving dossier would then form the core of the knowledge base that could be further augmented by domain experts during team visits to the engine and aircraft manufacturer's facilities.

The initial knowledge base was formed from an extensive review of the similarities and differences between ECU product lines, for business jet and large aircraft applications. Aero Engine Controls also defined typical usage patterns with associated environmental stresses, primarily based upon their experience with large aircraft systems.

3.1 ECU Product Line Review

A thorough review of the ECUs used on large airliners and business jets was carried out by Aero Engine Controls. It was concluded that there was very little, if any, difference in design processes, electronic component selection criteria, manufacturing methods and design verification techniques. In many cases, the ECUs were manufactured from identical component parts.

3.2 Environment and Operational Usage

While the operation of large civil aircraft does vary somewhat between airlines, most will follow a similar usage pattern. Aircraft will be scheduled to fulfil a series of flight legs, for a period of days, before time is allocated for maintenance activity. Aircraft will tend to be used on daily basis, with an average stage length of approximately four hours and a daily usage of up to eight hours. Some operators may use aircraft to carry out two flight legs per day allowing despatch and return to a main hub on a daily basis. Others may opt for a circular network of flight legs which will return to the main hub after a number of days.

By contrast, the majority of business jet operators do not operate their aircraft in a regular fashion, but schedule flights on an as-required basis, often no more than once per week. While a typical large civil aircraft may operate for 3000 hours in a calendar year, a business jet is unlikely to reach 500 hours.

However, there is a growing subset of business jet operators who offer their services to third party customers; this can have the effect of distorting the perceived average usage.

The ECU temperature profile is well understood by Aero Engine Controls from their experience of many large aircraft systems in operation today.

Engines will be started shortly before taxi, at which point the ECU temperature will be driven primarily by the ambient air conditions, with an element of self heating. Once the engine is started, the ECU temperature may rise due to conduction in the engine casing. Following take-off, the aircraft will ascend to a cruise altitude of 40,000 ft. In this flight condition, which will be held for the majority of the flight leg, the ECU Static Air Temperature [SAT] will decrease towards the outside air temperature of -40^{0}C. During descent and landing the ECU temperature will rise, once again, towards the ambient SAT, at ground level. Once the aircraft has landed and the engines are shutdown, the ECU temperature may increase due to engine soak-back but during this time the EEC will be de-powered.

Provided maintenance action to the engine control system is not necessary following the aircraft landing, the ECU will remain de-powered until the flight crew restart the engines, to taxi the aircraft for the next flight leg.

4. Exploratory Data Analysis [EDA]

The investigation team carried out a preliminary EDA to define the current baseline position with respect to reliability performance. This activity was scheduled prior to visiting the engine and airframe facilities in the hope that further questions would be uncovered that may prompt valuable discussion with domain experts. Three views of the reliability performance were captured using Pareto analysis, life data analysis and standard reliability metrics.

4.1 Preliminary Analysis

Pareto Analysis: Printed Circuit Boards [PCBs] were the only component type highlighted with significant, multiple, rejections. It was noted that PCBs of the same type are used in many other Aero Engine Controls products with no apparent reliability issue.

Life Data Analysis: Weibull analysis was carried out at the product [black box] level, using censor data generated from flight usage information. The analysis gave a reasonable fit to the data indicating a wear-out phenomenon ($\beta > 2$).

Standard Reliability Metrics: To complement the Mean Time Between Failure metric, the team also calculated the Mean Cycles To Failure. This analysis showed that if reliability was to be measured in flight cycles [MCBF] rather than flight hours [MTBF] then there would be no discernable difference in performance for an ECU fitted to either aircraft type.

4.2 Extracting Information from Analysis

While the confirmation of a long tail and a unit-level wear out mechanism were not unexpected, the impact of flight cycles upon reliability certainly was. The team decided this finding warranted further thought prior to discussions with engine and aircraft experts.

Comparing the flight profile for a business jet and a large aircraft, it can be seen that from the same starting position at take-off, the agile business jet will reach cruise altitudes much faster than the heavier civil airliner. Further to this, many business jets are certified for cruise operation up to 60,000ft and so will experience a lower minimum temperature.

The team postulated that a business jet may be subjected to flight cycles with an increased range and rate of change of temperature, when compared with a large civil aircraft. The business jet will also spend less time in the cruise condition where the temperature is most benign to component failure; figure 3 illustrates this comparison. If the business jet ECU were to experience increased cyclic stress, as described above, while operating a reduced flight cycle time, this would go a long way towards explaining the large difference in reliability, when expressed in terms of MTBF.

The team asked themselves the question, if the ECU reliability is being driven by flight cycles, then:

- ✈ Can this explain the elongated Pareto tail ?
- ✈ Are the observed failure mechanisms compatible with cyclic stress ?

These questions were deferred until after their fact finding visits.

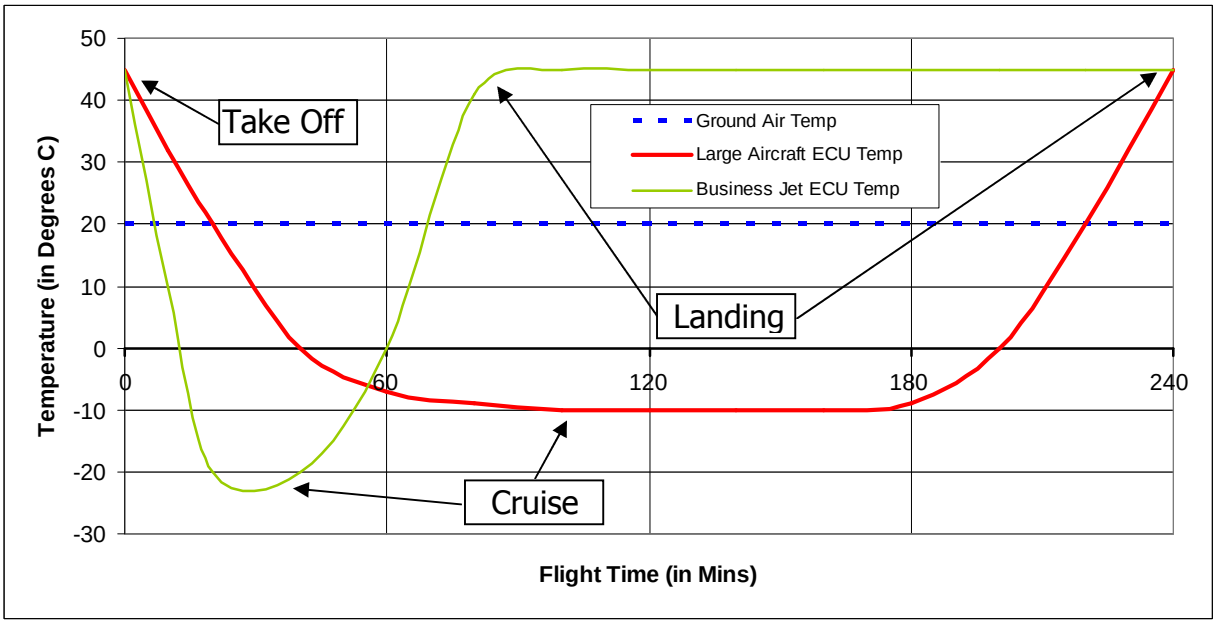


Figure 3: Flight Profile - Temperature Variation

4.3 Fact Finding Visits

Prior to visiting the engine and aircraft manufacturing facilities, the team arranged a short visit to the ECU repair facility to review the fault diagnostics process and to elicit product failure knowledge from technical staff. A system map of the repair and overhaul process was developed to determine the sequence of testing carried out to locate faulty parts within returned equipment. A combination of thermal cycling and vibration testing is used to re-create the environment in which the ECU is operated. This method has shown to be very effective with a high success rate in determining the root cause of component failure responsible for the system effect. Furthermore there was no evidence of a No Fault Found issue, suggesting that ECUs are being rejected from aircraft for legitimate reasons.

The nugget of knowledge taken from this visit was related to the completion of investigation paperwork. The diagnostic technician must identify the number of hours and cycles that the ECU has operated, at the time of its rejection, from information provided by the aircraft operator. Where this information was not made available, the technician would substitute with operational data, stored within the ECU memory. Following the visit, the effect of mixing these data sources was analysed and is described in section 5.1

A visit to the engine manufacturing facility was carried out to consolidate understanding of the stresses associated with the assembly process and to confirm assumptions regarding the engine operating environment. While an in-depth critique of the build and assembly process uncovered little risk to ECU reliability, a review of flight test data indicated a flaw in operating envelope assumption. Data generated from an ECU, specifically instrumented to monitor the thermal environment during flight test, indicated that the SAT did not follow the outside air temperature as expected. The maximum recorded temperature was higher than expected and the total range, less than 20^0C.

A visit to the aircraft manufacturing facility by the investigation team enabled discussion with a variety of experts to develop a clearer understanding of the operational stresses on the ECU during the life of a business jet. This provided a very effective means of eliciting critical pieces of information that was until this time, unavailable or not thought to be relevant. Following the visit, key facts were consolidated in the knowledge base.

- Unlike a large aircraft where power is removed following landing, an ECU fitted to a business jet may be powered at all times that the aircraft is powered.

- Business jets may remain powered for elongated periods of time waiting for an executive to complete their meeting.

- Following assembly and initial flight test, each new aircraft will spend many weeks at a completion centre where it will be fitted with specific customer furnishings; during this period the aircraft may be powered continually.

5. Data Analysis

The information accrued during the task team visits was reviewed extensively to validate, discount or add to the information captured in the knowledge base. It was found that a number of the findings had altered or discredited some of the assumptions on which the initial analysis has been based.

5.1 Consolidation – Revisiting Initial Assumptions

An important distinction between sources of operating time used to determine the ECU reliability performance was uncovered during the visit to the ECU repair facility. At the aircraft level, the flight time is used to define ECU usage while at the engine level it is the time that engines have been running. These measures are, in practice, very similar and would have little bearing upon the accuracy of reliability monitoring. However, when this information is not made available on the rejection paperwork, the repair facility will record the hours that ECU has been powered, from data stored within the ECU itself.

This measure of ECU usage will be significantly larger than the engine running or aircraft flight time as the ECU is now known to be powered for many hours between flights. As the current service data is known to have a mixture of these usage figures, the MTBF metric can no longer be seen as a representative measure and the Weibull analysis must be called into question.

It became apparent during the visit to the aircraft manufacturer that the ECU may be powered for a significant number of hours between flight legs, increasing the usage time considerably when compared with the length of the flight leg. Furthermore, the ECU may spend many weeks powered constantly while the aircraft is furnished within the completion centre.

The selection of usage data prompted the question of which source would be most appropriate for reliability monitoring. While the use of aircraft flight hours provides a performance metric that reflects user perception, it does not capture the total time the ECU is powered, which may be a reliability driver.

Additionally it is also now clear that the ECU operates over a much smaller temperature range than was initially assumed, far smaller than an equivalent ECU installed on a large aircraft. This brings into question the assumption that the reliability may be driven by flight cycles.

The data provided from flight test is of greatest significance, indicating that the ECU is at its highest temperature while powered, with the aircraft on the ground and with the engines not running. For the business jet, unlike the large airliners, this high stress condition is where the ECU will spend most of its powered life.

In summary, data obtained from experts during the task team visits has discredited the cyclic failure assumption and has suggested that the time at high temperature may be the key reliability driver, differentiating business jets from large civil airliners.

5.2 Modelling Reality

The team developed a plan to validate the revised assumption that it is powered time at elevated temperature that drives ECU failure rate. In order to model this scenario, two data sets were required; *time to failure* for each rejected ECU and *censor data* for the entire business jet fleet. Time to failure data was made available by the ECU repair facility but the fleet censor data presented a problem.

The measure of ECU usage at the aircraft level, where the data is recorded, is in *flight time*, while the time to failure data is defined in ECU *powered time*. To compound this issue there is also a significant period of time unaccounted for, where the ECU is powered while the aircraft is in the completion centre.

The team recognised the only credible way to carry out this analysis was to generate a new dataset which represents fleet usage, but defined in ECU powered hours. In order to create this dataset, the team needed to define the relationship between flight hours and ECU powered hours. To accomplish this, a sample of ECU rejection data, for which both ECU powered hours and fleet usage hours were available, was examined. A simple comparison of ECU powered hours against aircraft flight hours, using 32 data points, showed a good visual correlation; this is shown in Figure 4.

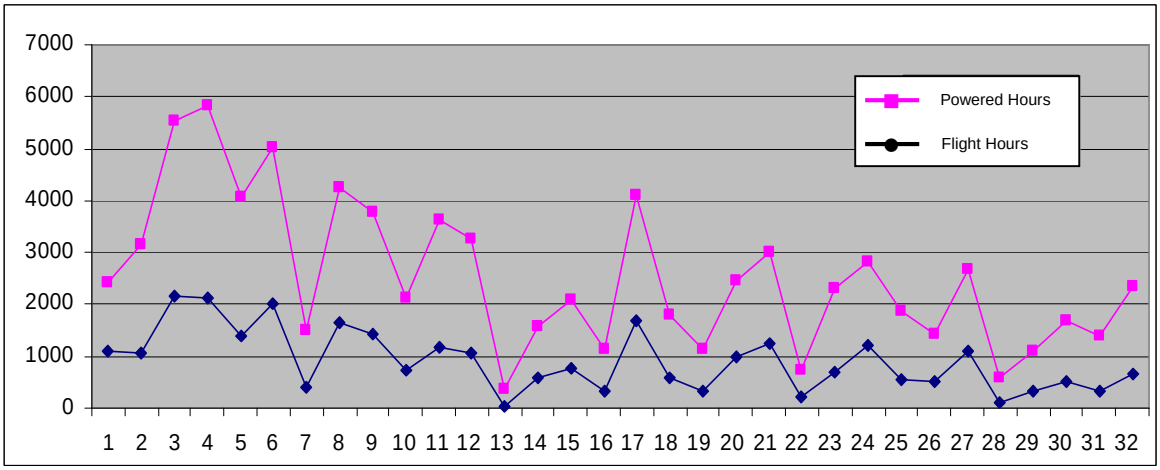


Figure 4: Comparing ECU Powered hours with Flight Hours

Formal correlation testing displays a remarkably close relationship between aircraft flight hours and ECU powered hours, see figure 5. The line of best fit indicates that an ECU is powered for a period 2.5 times the accrued flight time, with an offset of approximately 400 hours.

The team were comfortable that the emergent relationship could be justified in practical terms. The factor of 2.5 represents the time powered between flight legs and the 400 hours offset, offers a reasonable approximation of the time powered in the completion centre.

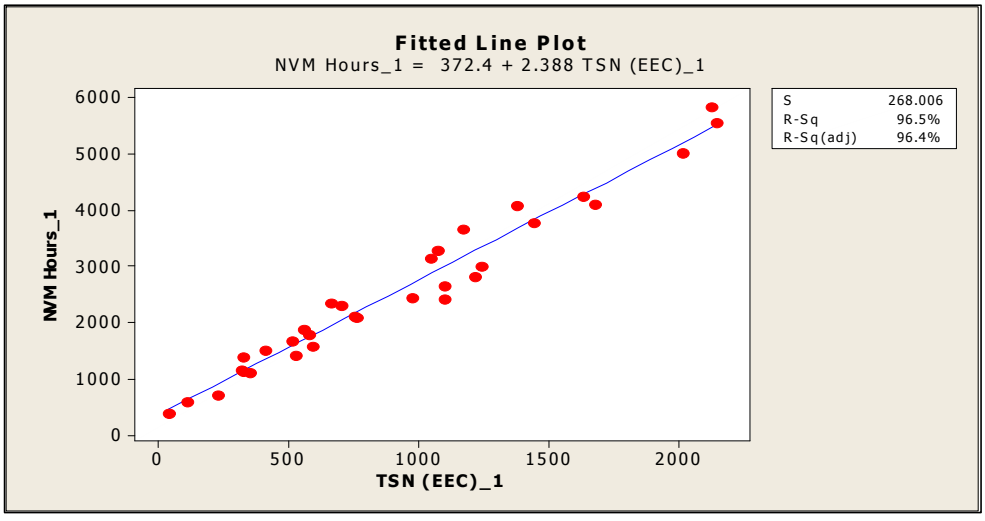


Figure 5: Statistical Correlation Testing

6.0 Searching for Failure Mechanisms

Revised Weibull analyses were shown to indicate an excellent fit to the newly generated dataset. The high confidence afforded by this data prompted the team to focus their efforts on life data analysis for each dominant component failure, as defined by the Pareto. Using the analyses as a guide, the team set out to locate the underlying physical processes that could drive the observed failure characteristics.

As an example, revised life data analysis for PCB rejections highlighted a strong wear out mechanism. This finding was in direct conflict with the original assumption that suggested that thermal cycles were the cause of PCB infant mortality failure. Armed with the revised failure distribution as a guide, the team carried out a series of focused inspections on the rejected PCBs. The team found evidence of contamination within the PCB structure which, when powered at raised temperature, would gave rise to electro-migration. It was shown by test that this process would form a conductive path between PCB tracks, leading to failure. Testing carried out on a sample of unused PCBs confirmed a similar failure characteristic to that exhibited by the PCBs removed from the service environment.

Following a brief from Aero Engine Controls, the PCB supplier found traces of the contaminant in their manufacturing environment and took steps to ensure its removal. There has been no evidence of PCB failures since this process change.

Capacitor and diode failures were investigated using similar methods. In both cases, accelerated test confirmed the presence of a latent failure mechanism consistent with component failure in the service environment. Alternative components were selected and tested, in an equivalent manner to the original devices, to verify the anticipated improvement in performance.

The resultant changes to the design of the ECU, described above, have improved its service reliability considerably.

7.0 Conclusion

The investigation team has initiated a number of ECU design changes to improve reliability, each closely related to high temperature operation.

While the reliability did improve dramatically as the design changes were incorporated, improvement action has not been possible for all component failures. This is mainly due to the small sample of failures for many of the components, making correlation with analysis very difficult. Further work is required to better understand and characterise the thermal profile of the business jet, particularly in the areas local to the ECU installation. This information will then provide further guidance and drive focus in understanding the failure processes at work in the remaining components.

An investigation strategy has been defined which promotes a team approach for gaining insight into component failure. The combined, iterative, use of data analysis, expert knowledge elicitation and service observation has proved a very powerful process.

Following this initiative, the ECU has shown dramatic improvement in its reliability performance and the team have certainly made progress towards understanding the generic factors that influence the reliability of ECUs used on business jet applications. It is possibly just as important that the team have, along the way, discredited some commonly held beliefs regarding business jet operation which can now be seen to have thwarted previous attempts at reliability improvement in this area.

References

[1] Patrick O'Conner, *Practical Reliability Engineering*, Wiley, 2002

[2] CALCE/EPSC, *Physics of failure in Reliability Assessment*, CALCE Workshop, Farnborough, 1999

[3] Wayne B Nelson, *Accelerated Testing Statistical Models*, Test Plans and Data Analysis", Wiley, 2004

[4] Robert Abernethy, *The New Weibull Handbook Fifth Edition*, 2006

[5] Wayne B Nelson, *Applied Life Data Analysis*, Wiley 1982

The Dependability Case Is It Achievable

Richard Denning, Nick Barnett

Ministry of Defence Abbey Wood – South, Bristol BS34 8JH

Abstract

The Ministry of Defence (MOD) has been using the concept of the Reliability and Maintainability (R&M) Case as part of it's method for achieving dependable systems for over 10 years. The in-house standard for the R&M case (Defence Standard 00-42 pt 3) has recently been used as a basis for developing the new British Standard BS5760 pt 18 – Guide to the demonstration of dependability requirements – The Dependability Case, which in turn is being used as the starting point for IEC 62741 – The Dependability Case.

The R&M case, and hence the dependability case, focuses on identifying the risks associated with the achievement of R&M / dependability then undertaking mitigating activity and making an argument that the R&M / dependability of the product will be achieved or is being achieved (if it is past the in-service point)

Experience with the R&M case approach has identified a number of issues with regards developing programmes of mitigation and then making a sound argument that the product will be / is reliable and maintainable. These issues are still a regular occurrence 10 years after the introduction of the methodology and are likely to be compounded by the expansion to dependability as dependability is a much broader subject than traditional reliability and maintainability.

This paper discusses typical poor use of the output of individual R&M activities to argue that a project will/has good R&M characteristics and how with relatively small additional effort the output could be used to better effect.

1 Introduction

In late 1999 the MOD introduced the concept of the R& M Case, as a tool to aid it to gain confidence that the R&M characteristics of the systems it was buying would be fit for purpose. The approach was introduced following discussion with industry who felt that being asked to contract to a shopping list of reliability activities was not the way to either deliver a reliable and maintainable product or to drive a programme to be cost effective and timely.

1.1 Standards for R&M Case

The R&M case is described in detail in Defence standard 00-42 part 3,[1] the approach has been incorporated into the UK national standard BSI 5760 PT 18[2] and is now being developed as international standard IEC 62741.[3] There is also an American SAE standard[4] which uses the same general approach and this is called up by the NATO Publication ARMP1[5] which is the standard

for use when procuring a NATO project. In line with current NATO policy ARMP 1 is currently being updated to invoke IEC 60300-1 & 2 and IEC 62741

1.2 Summary of the R&M Case approach

The R&M case approach is based around 3 objectives:

1. The R&M requirements of the purchaser shall be determined and demonstrated to be understood by the purchaser and contractor

2. A programme of activities shall be planned and implemented to satisfy the requirements and investigate the risks identified

3. The purchaser shall be provided with progressive assurance that the R&M requirements are being, or will be, satisfied and that confidence is building

Objective 1 requires a dialog between the purchaser and contractor to ensure both understand what is required. This dialog should be initiated by the purchaser raising an initial R&M case which details what the requirement is, why this requirement is needed and why it is considered to be feasible.

In order to meet objective 2, it is assumed that the contractor knows what activities need to be undertaken to mitigate the R&M risks that have been identified to ensure the delivery of reliable and maintainable equipment and can articulate what they are going to do and how it adds value. Objective 3 is satisfied through the R&M case which is defined as a reasoned, auditable argument created to support the contention that a defined system satisfies the R&M requirements.

The approach also moved from a programme built on a list of tasks (often specified by the purchaser) to a risk based approach where the supplier would assess the product and understand the risks associated with achieving the required levels of R&M performance. The Supplier would then design a programme of activities to mitigate the risks and ensure that the product achieved the required levels of performance.

1.3 Issues with the case approach

From the description of the R&M case it can be seen that compared to a more traditional approach (contracting for a list of R&M related activities, which when completed completes the R&M programme) there are a few extra stages to the process:

- Using risk assessment to decide on the programme of activities required to give all stakeholders confidence that the product will be reliable and maintainable.

- After completing a given activity, considering what it means and updating the argument that the final product will have appropriate R&M characteristics.

- Assessing how the output of an activity has increased the knowledge of the product's R&M characteristics and how this new knowledge impacts the risk and hence the programme of activities.

The first and last stages are reasonably similar and have suffered around a number of issues, typically:

- Understanding what the risks are
- Understanding what sort of activities will help to mitigate the risks
- Justifying the size and scope of selected tasks
- Making the business case to spend time and money on activities

These issues have been discussed in some detail in a number of papers[6,7,8,9] therefore this paper will address the issues associated with making best use of the results of R&M activity to influence design and structuring the argument to give the purchaser confidence that the product will be reliable. Improving the use of the results from R&M activities, will improve the confidence the purchaser has in the product and should reduce conflict and reduce nugatory work.

2 Making the case

2.1 Problems with making the case

Having completed any activity there is a need to take the output from the activity (Evidence) and apply some logic to make an argument that the product will be good.

So after a reliability test which the product *passed*, it is not acceptable for a statement that the item has passed the test or more often a reference to a test report. The minimum required is a statement to show:

- The sample tested was representative
- The test is a suitable representation of expected life
- Appropriate analysis has been applied to justify that the test has been passed.

Unfortunately a large proportion of R&M cases fail to make a convincing argument, although with slightly more thought and intellectual effort many R&M cases could be considerably improved as will be shown in the remainder of this paper.

2.2 Evidence from A Maintainability Demonstration

A system had under gone a comprehensive maintainability demonstration, at the end of which it was concluded that a sub-system Mean Active Repair

Time (MART) was 10 minutes longer than the contracted requirement. This was included as a statement in the R&M case as follows:

…. "The maintainability demonstration on the remote sensing sub system was witnessed at the premises of our sub-contractor with staff from both companies present. A MART requirement of 15 minutes was flowed down to them but the results of the demonstration show a recorded MART figure of 25.45 minutes, thus the test has been recorded as a failure."

This immediately resulted in a large discussion and much effort on behalf of the purchaser, where it was discovered that the extended time was being driven by a couple of repair actions in the tail of the distribution which were rare activities and which the company thought there was a potential to improve. This resulted in a reissue of the R&M case:

…. "The maintainability demonstration on the remote sensing sub system was witnessed at the premises of our sub-contractor with staff from both companies present. A MART requirement of 15 minutes was flowed down to them but the results of the demonstration show a recorded MART figure of 21.45 minutes, thus the test has been recorded as a failure.

Investigation has shown that the Mean Time is being adversely affected by two tasks that require the access plate to be removed and resealed, the cure time of sealant being 1 hour. The cure time was not allowed for in the original maintainability predictions.

We are investigating alternative sealants to find one of the same specification but with a shorter cure time. We are also reviewing the access route needed to remove and replace the widget to see if it can be done without the need for removing the access plate.

If neither investigation provides a suitable alternative we do not believe the extended Mean Time will cause significant issues in service as the failure rate of the items concerned is such that we do not anticipate needing to replace them more than once every 7 years as similar items used in a similar way have a demonstrated life in excess of 7 years."

This new R&M case made it clear that although currently the system was failing to meet the requirement, the supplier had options to improve the repair time and even if this was not possible, then the impact on the system performance was understood and it was not considered to be major.

2.3 An example of a FMECA ISSUE

During the development of a major system, a detailed FMECA was produced which discovered an issue with part of the system, which was reported in the R&M case as

…. "The FMECA has identified a failure mode where secure communication could be compromised as there is no obvious indication to the operator that

the encryption system has gone off line. The operator will need to leave his station and walk round to the encryption tray to check that the red indicator is not illuminated prior to any communication activity."

This was identified by the purchaser as unacceptable and when pointed out to senior management that this was the case they were genuinely surprised that their team could think it was acceptable. This was a case of an activity being done, but the impact of the findings not being considered and / or acted on. It would have been better to do the activity and think about the findings which could have resulted in a Case statement

…. "The FMECA has identified a failure mode where secure communication could be compromised as there is no obvious indication to the operator that the encryption system has gone off line. Without modification the operator will need to leave his station and walk round to the encryption tray to check that the red indicator is not illuminated prior to any communication activity.

We acknowledge that having to leave the station prior to any transmission will adversely affect system operation and are working with the supplier of the encryption system to identify ways in which the indicator can be 'repeated' or repositioned elsewhere in the system such that the operator can check the status of the encryption system from his / her station. Ideally this will be through a change to the BIT / BITE system giving an indication on the main status page but if this is not possible we have identified a number of options for making the red indicator visible from the station.

The Project and R&M plans have been updated to show this additional work as until the issue is resolved it is considered to be on the critical path for the project."

This would have demonstrated that the supplier was acting in accordance with the concept of the R&M case

2.4 Reliability Modelling Issue

Very early in the development of a complex multifunction system, an RBD was developed to predict the R&M characteristics of the system. The statement in the R&M case report was:

…."A Reliability Block Diagram of the sensing sub-system has been constructed that results in a prediction which shows that this system will be 500 hours short of the required MTBF.

A Reliability Block Diagram of the data processing sub system has been constructed that results in a prediction which shows that this system will be 150 hours short of the required MTBF

A Reliability Block Diagram of the display sub system has been constructed that results in a prediction which shows that this system will exceed the required MTBF by 1000 hours."

This highlighted that the system would not achieve the requirement – so the R&M case was not giving the purchaser confidence that the system being developed would meet the needs of the purchaser. It also implied that the supplier did not use reliability techniques to drive the design and the only point of carrying out analysis of the reliability was to assess what would be delivered.

Following robust discussion the programme of activity and approach was changed and hence the case was updated to:

…."We have constructed Reliability Block Diagrams of the three sub-systems with results showing that 2 are currently failing to meet their required levels whilst 1is exceeding it.

Experience from previous projects is that predictions done at this early stage of the programme are often pessimistic as the data is immature, significant amounts of it coming from commercial databases or supplier information. On the previous 2 programmes we found that the levels of reliability achieved by the end of design exceed the predictions done at a corresponding stage by up to 40%.

We have reviewed each of the models and in one area are recommending a design change to increase redundancy as a sensitivity analysis suggested that even a 10 fold increase in reliability of widget A would not improve the overall reliability of the sub-system by more than a few percent.

Data from a similar project which has recently been fielded for another customer is becoming available through our in house DRACAS system and is already showing that some of our derating assumptions based on environmental factors were harsh and we envisage that by the middle of this phase we will be able to submit an updated modelling report based on better data and incorporating the design change referred to above that will give more confidence that the overall system will be capable of meeting its operational requirement.

The RBD's and all of the associated data are reproduced in Annex A to this Case Report for completeness"

This clearly showed that R&M was impacting the design, rather than being a bolted on after thought and helped to regain the purchaser's confidence that the programme would deliver a system which met it needs, although the purchaser's confidence was weaker than it would have been if there had not been the need for additional dialogue after the original delivery of the R&M case report.

2.5 Testing Issues

Following detailed testing of a prototype vehicle the R&M case contained a lot of detail about early R&M modelling, FMECA activity etc, but very little detail on the testing which had been undertaken, this being covered by:

…."We have undertaken a series of demanding tests on the vehicle completing many battle missions, during which time we have recorded incidents as they occurred, all of which have since been investigated and sentenced in accordance with the standard. The results of this trialling show that the vehicle has exceeded its design requirements."

This made it appear that the supplier was relying on predictive techniques rather than test results which implied that they did not have confidence in their test results. This was not the case, and following discussion the supplier updated their report to:

…."The vehicle completed 200 Battlefield Missions (BFM) during which time 10 mission failures were recorded along with an additional 42 basic failures. The BFM was 24 hours in duration with 60% of the travelled distance (150Km) being off road. The vehicle spent 2 periods of time, totalling 4 hours, parked up with the engine running whilst the internal equipment was in 'listening' mode. For 2 days of the trial the temperature was in excess of 25 degrees centigrade requiring the air conditioning to be running for long periods of time. The full details of the trial, a spreadsheet listing all of the incidents and how they were sentenced and the final trial report are all referenced within this case report and available for review as required.

This clearly showed that the R&M activity was being used to give confidence that the final product would meet the requirements.

3 Conclusions

The R&M case approach means that people can not just do R&M activities; there is a need to think about the results and act on these results. Even those undertaking a sensible programme of activities and acting on the findings can give poor confidence due to not explaining how the R&M programme is impacting the development of the physical product.

Those organisations who have seen R&M activity as an overhead which is bolted on to the design and development process struggle with the R&M case approach as often their R&M activity is too late to influence the product and all the case can do is document how poor the product is likely to be.

A number of suppliers undertake prediction activity and when the modelling results show that the product will not meet reliability requirements, state that in their experience predictions always under estimate the actual reliability. But when challenged for evidence can not produce anything.

Unsupported statements and claims in any assurance case, whether positive or negative, do not give the reader any confidence that the work undertaken has added value or mitigated any of the identified risks.

The use of the R&M case is (after 10+ years) starting to move people away from thinking that the completion of an R&M activity or the production of an R&M case is the goal of the R&M programme and moving people to see that

the goal of an R&M programme is to influence the physical product and ensure that the R&M requirements are met.

With some wider thinking about what the outputs from a R&M activity really mean and a willingness to adjust the direction of the R&M programme the Dependability Case can demonstrate that a product will be / has achieved the desired levels of R&M.

References

1. Ministry of Defence, Defence standard 00-42 part 3, Issue 4, Reliability & Maintainability Assurance Guides. Part 3 - R&M Case

2. British Standards Institute, BS 5760-18:2010 Reliability of systems, equipment and components – Guide to the demonstration of dependability requirements – The dependability case
3. International Electro-technical Commission - IEC 62741 -The Dependability case
4. SAE
5. NATO – Allied Reliability and Maintainability Publication - ARMP-1: NATO Requirements for Reliability and Maintainability
6. SaRS Journal Vol 22 No 2 – R&M Case Edition (2002)
7. Denning, 3 years experience of the R&M case - ESREL 2003, pp 489-493
8. The R&M Case – How the MOD Assures itself that it is getting what it needs - SaRSS 2005
9. N Barnett The R&M Case – A Decade On, SRE seminar RAM V Workshop - Bringing RAM Value In a Declining Resource Environment.

Onboard, Real-Time Detection of Adhesion Levels in the Rail/Wheel Contact

Peter Hubbard, Chris Ward, Roger Dixon, Roger Goodall
School of Electronic, Electrical and Systems Engineering,
Loughborough University, Loughborough, UK

Abstract

Low adhesion in the wheel/rail contact or the 'leaves on the line' problem is a large operational issue for the railway industry. There is currently a shortage of up to date information about the running conditions of rails with respect to short term adhesion trends (over a daily period) and macro trends (across seasons). This can lead to costly over application of mitigation actions such as rail head cleaning to combat the problem.

The generally established methods of assessing areas of low adhesion involve mapping activations of wheel slide and wheel slip protection events to track locations. These methods are reactive and rely upon a slip/slide event to be initiated by the application of traction or braking. Research presented here is part of an RSSB managed project that forwards previous fundamental research into methods of Low Adhesion Detection (LAD) in real-time. These techniques are based around using 'modest cost' inertial sensors mounted to in service vehicles. The LAD system proposes that the motions of a railway vehicle (in both lateral and yaw movements) vary as the adhesion conditions under the vehicle change. If the changes in the running dynamics as a result of low adhesion can be observed and interpreted, they can infer the adhesion at all points across a network and not just when slip/slide events are triggered. These approaches have been verified against rail vehicle simulations performed by DeltaRail using the multi-body physics software VAMPIRE®.

1. Introduction

Across the rail network, low adhesion problems caused by track contaminants is a complex problem for the rail industry to manage. Levels of adhesion are subject to variation over the short term (across as a daily period) or over longer terms (effects due to seasonal change). The 'live' measurement of adhesion in the wheel rail interface of a vehicle operating under normal conditions is currently not available. Established methods of identifying these problem areas involve mapping wheel slide or wheel slip events to track locations. This leads to difficulty assessing the current operating risk of rail vehicles and results in the use of 'catch-all' mitigation solutions; such as the over application of railhead cleaning or the pessimistic rescheduling of operations based on a global forecast.

The research presented here is part of the RSSB managed project T959, a follow on project from the feasibility study T614. These projects are concerned with methods to detect areas of low adhesion under normal operating conditions using 'modest cost' inertial sensors mounted to service vehicles. The goal of such a system is that estimations taken will better inform risk

mitigation activities across the network. This will enable the improvement of scheduling management under varying conditions and reduce the cost of maintenance operations (such as railhead cleaning) by targeting them at problem areas.

Techniques proposed in [1–3] suggest that the motion of a railway vehicle (particularly in the yaw and lateral motions) varies as the adhesion levels change. If these changes in dynamics can be observed under normal operating conditions and attributed to the contact forces, an assessment of the level of adhesion currently experienced can be made.

This paper focusses on two of the previously suggested techniques [1–3]; a model-base estimator approach and a direct-data analysis approach. Both of these techniques rely on capturing the appropriate vehicle dynamics by inertial sensors attached to wheelsets, bogies and the vehicle body. The model-based approach approximates adhesion by first deriving the level of creep force creep force (i.e. the contact reaction forces) between the wheel and the rail as shown in [1]. It has recently been found [4] that contrary to initial suggestions [5] the initial slope on the relationship between ‘slip’ and creep force varies with the level of adhesion. This means that the level of adhesion can be derived from the overall values of creep forces should an approximation for the amount of slip be found. Further explanation of this method and the resultant findings are presented in section 2.

The second method progresses from a solution presented [6] where the adhesion level is approximated by analysing how the relationship between different vehicle dynamics change as adhesion changes. This study, on data from an older style vehicle, showed that different parts of the vehicle moved in increasingly similar or dissimilar ways as adhesion levels changed. By quantifying the level of correlation of particular dynamics over a sample period, it was possible to approximate the level of adhesion experienced. Section 3 presents the results of this study.

This project was driven by a committee composing of academic and industrial partners from the rail industry. This allowed the research team to understand the operating context of an adhesion detection system from the point of view of industrial experts. Section 4 presents some ideas captured from these discussions about where the current capability of the system would be operationally beneficial and where there is opportunity for development to provide advanced capabilities.

2. Model-Based Estimator

The technique of using a model-based approach to estimate contact forces under normal running conditions has been proven against linear suspension models in a MATLAB/Simulink environment [1–3]. Part of the progression of the work within this project is the verification of creep-force estimation using simulation data taken from VAMPIRE®. This data is treated as if it collected in real time from an in-service vehicle in order to provide a suitable level of validation of the capability of the creep force estimator as in [6].

2.1 Methodology

The model-based estimation concept used in this study is based around the use of the well-known Kalman-Bucy filter [7]. In order to form this filter, a linear, mathematical model describing the suspension system dynamics is required to be derived. Previous studies [8], [9] have shown that the dominant dynamic characteristics as a result of track irregularity are contained only in the lateral and yaw planes of motion. As such, the linear model derived need only be concerned with the description of suspension in this 'plan-view' sense.

As mentioned previously, the primary problem in this application is that creep forces cannot be measured directly. The solution employed in this case is to manipulate the Kalman filter to output the creep forces as 'augmented states'. In this method, the numerical values of these states are defined by being attributed to the 'left-over' values from the force/balance equation. For this method to be successful, an accurate linear model is required as any processing noise will be interpreted as additional creep forces.

2.2 Model Development and Verification

The vehicle used as a case study for this work is a generic modern passenger vehicle with characteristics based loosely on the British Rail class 158 design. This particular suspension design is sympathetic to this estimator concept as it has largely linear characteristics and will map well to the chosen modelling technique. In addition, this contemporary design should mean that the results found here will be applicable to a wide range of modern vehicles. Figure 1 shows a schematic of the linear adaptation of the primary suspension of this vehicle.

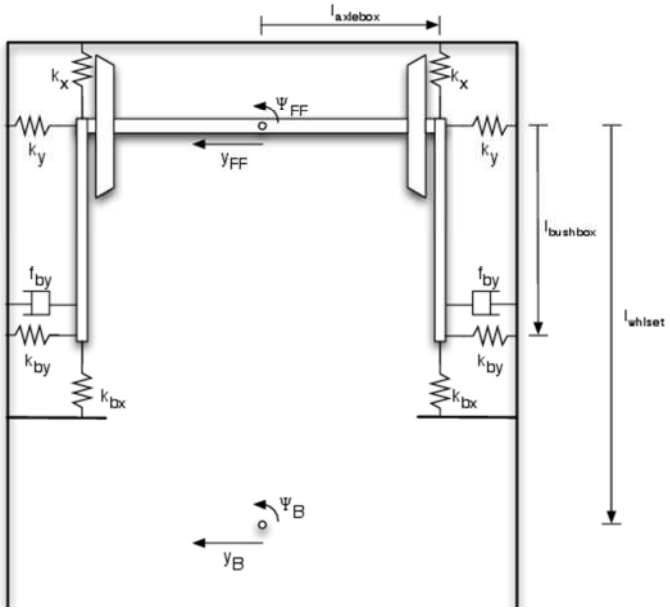


Figure 1. A suspension schematic for a generic modern passenger vehicle bogie

In Figure 1 the following describes the parameters labelled:

- $k_{x,y}$ – Longitudinal and lateral primary spring stiffnesses
- $k_{bx,by}$ – Longitudinal and lateral bush stiffnesses
- $f_{bx,by}$ – Longitudinal and lateral bush damping
- $l_{component}$ – moment arms
- y_{FF}, ψ_{FF} – lateral and yaw position of front wheelset on the front bogie
- y_{FF}, ψ_{B} – lateral and yaw position of front bogie

The aim here is to derive a description of the dynamics of the system in standard state space form, which is defined in equation 1.

$$\dot{\boldsymbol{X}} = \boldsymbol{A_k X} + \boldsymbol{B_k U} \tag{1}$$

If the states and inputs are defined in the following way:

$$\boldsymbol{X} = \left[y_{FF}\ \dot{y}_{FF}\ \psi_{FF}\ \dot{\psi}_{FF}\ F_{FF}\ M_{FF}\right]^T \tag{2}$$

$$\boldsymbol{U} = \left[y_B\ \dot{y}_B\ \psi_B\ \dot{\psi}_B\right]^T \tag{3}$$

The A_k and B_k matrices can be derived by forming the force/moment balance equations based on the parameters shown in Figure 1.

As described in detail by previous work done [2] the Kalman-Bucy filter is tuned primarily via the 'Q' matrix which identifies a degree of certainty with each of the state models. By setting the contact force state models as highly uncertain compared to the vehicle dynamics state models, the filter can be used to approximate the creep forces in real time. The values chosen are shown in equation 4.

$$\boldsymbol{Q} = \left[1e^{-6}\ 1e^{-3}\ 1e^{-6}\ 1e^{-3}\ 1e^{12}\ 1e^{12}\right]^T \tag{4}$$

The creep force output is verified by the used of recorded data from the multi-body simulation package VAMPIRE®. This software allows the accurate three dimensional simulation of a rail vehicle, inclusive of non-linearities in the suspension and wheel/rail contact. It also provides data outputs capturing all aspects of the vehicle such as the measurements of dynamics of the wheelset, bogie and vehicle body, and the values of creep forces experienced.

In the following test, a simulation was performed in VAMPIRE® where an equivalent modern passenger vehicle to the one depicted in Figure 1 was subject to a sixty second transit along a section of track. The vehicle travelled at line speed (in this case 200kph) and the track contained a representative amount of irregularities. In this test the level of adhesion on the track was set to a constant, ideal value. The dynamic measurements captured in VAMPIRE® were input into the estimator as if they were real-time

measurements. Figure 2 shows the comparison between estimated and recorded creep force and moment for a short time section of this run.

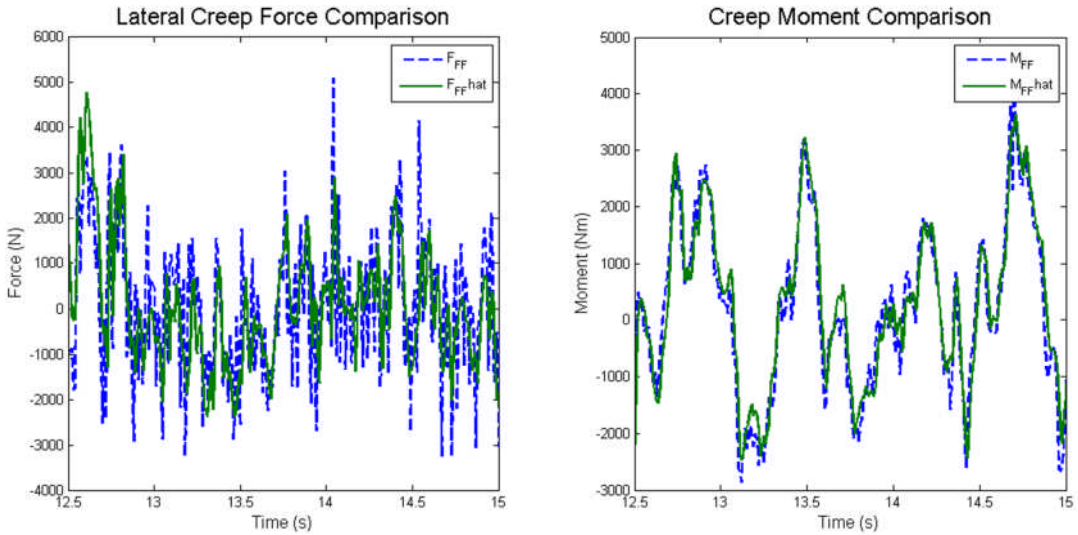


Figure 2. Comparison between recorded and estimated creep force (F_{FF} to $F_{FF}hat$) and recorded and estimated creep moment (M_{FF} to $M_{FF}hat$)

It can be seen that in this test scenario, the estimator performs well, particularly with creep moment. The errors observed are due to the small inaccuracies inherent to the simplification of the real vehicle to a linear, plan-view representation.

2.3 Post Processing Creep Forces for Adhesion Estimation

The method of using augmented states has allowed a real-time numerical output for a value that cannot be directly measured. In this case study, the value returned is that of creep force. Although this is a useful parameter to know, it does not directly infer a value of adhesion. The adhesion level can only be truly assessed if the creep force readings are evaluated against the amount of slip (i.e. differential velocity between wheel and rail) experienced.

If the vehicle travels along a perfectly smooth track, the total values of contact force in the lateral and yaw sense would be zero. It is the presence of track irregularities that agitate the wheel and cause a dynamic response of the vehicle. It can be assumed that the amount of slip generated must have some direct relationship with the track irregularity. Because both slip and track irregularities are unable to be directly observed from on-board the vehicle, they must be inferred from another measurement.

Figure 3 shows a moving 5 second RMS window applied to the creep moment values when considered along a sixty second run. It can be seen that in general the overall level of creep moment estimated falls as the adhesion level reduces. The change is verified by the comparison made with the creep moment as recorded by the VAMPIRE® simulation.

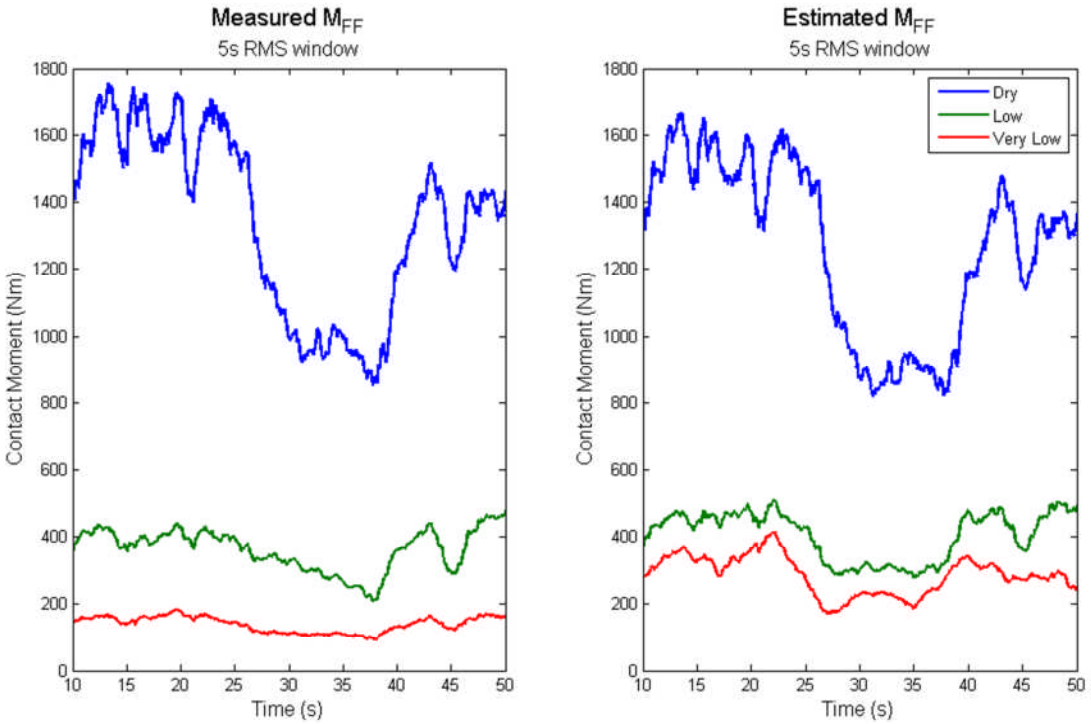


Figure 3. Comparison of Creep Moment (RMS) between VAMPIRE® recorded values and Estimated Values across 3 adhesion conditions.

The variation in creep moment as a result of the changing track irregularity distribution can be observed in all runs. The solution adopted here was to scale the creep moment RMS values with each of the on-board dynamic measurements, when an RMS of these is taken over the same period. It was found that scaling the result by yaw acceleration provided a reasonable solution.

In each of the three test runs shown in Figure 3, the average value for the ratio between estimated creep force (RMS) and yaw acceleration (RMS) is used as a calibration value for the associated level of adhesion. Therefore, if the same ratio can be found in a different scenario, the captured data can be used via a linear interpolation method to approximate a level of adhesion.

2.4 Verification of Model Based Approach against VAMPIRE®

Figure 4 shows two test cases that were used to validate the above method. In this test identical conditions are used in that in the previous section. In this test a step change in adhesion occurred after 30 seconds, highlighted by the 'dashed' lines that show the actual adhesion level experienced. It can be seen that the estimator correctly observes this change in adhesion and is able to distinguish between low and very low conditions. This is a significant observation as these two adhesion levels present significant different operations risks.

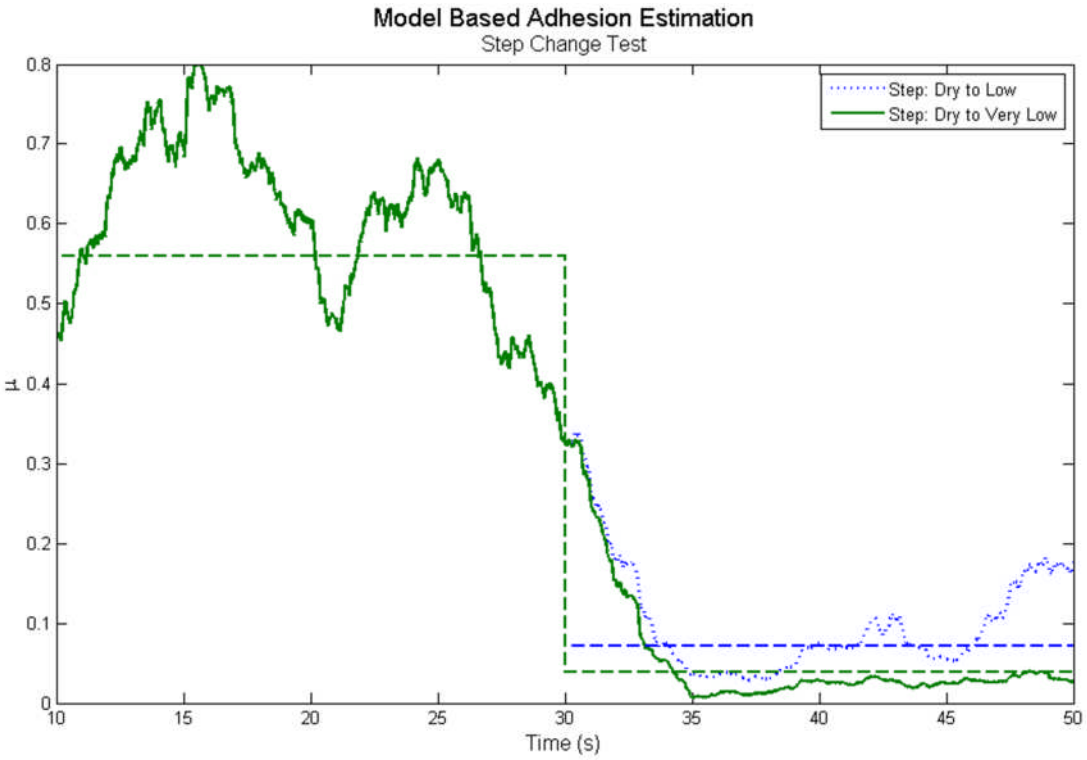


Figure 4. Figure showing model-based adhesion estimation for two step changes in adhesion.

Further verification of the model-based estimator is to occur by blind data tests, whereby no contact force data is provided and an estimation will be proven to have been made by running dynamics only.

3. Direct Data method for Adhesion Estimation

Alongside research into the model-based approach, analysis has been performed by observing the VAMPIRE® data outputs directly. The goal of this research is attempt to observe any changing characteristics of the dynamic data as adhesion level changed. It was observed previously [6] that this was true in the case of a Mk3 bogie model, but the same results was not observed using the generic modern passenger vehicle.

3.1 Correlation Analysis

By taking the VAMPIRE® data sets produced for the extreme adhesion conditions (namely ‘dry’ and ‘very low’), a sample cross correlation was performed for the leading and trailing bogie dynamics. This assessment was performed when considering a varying sample delay on the leading bogie set. This is to attempt to observe any increase or decrease in correlation as the trailing bogie passed over the same place in the track. The results of this analysis when performed on the yaw dynamics is shown in Figure 5.

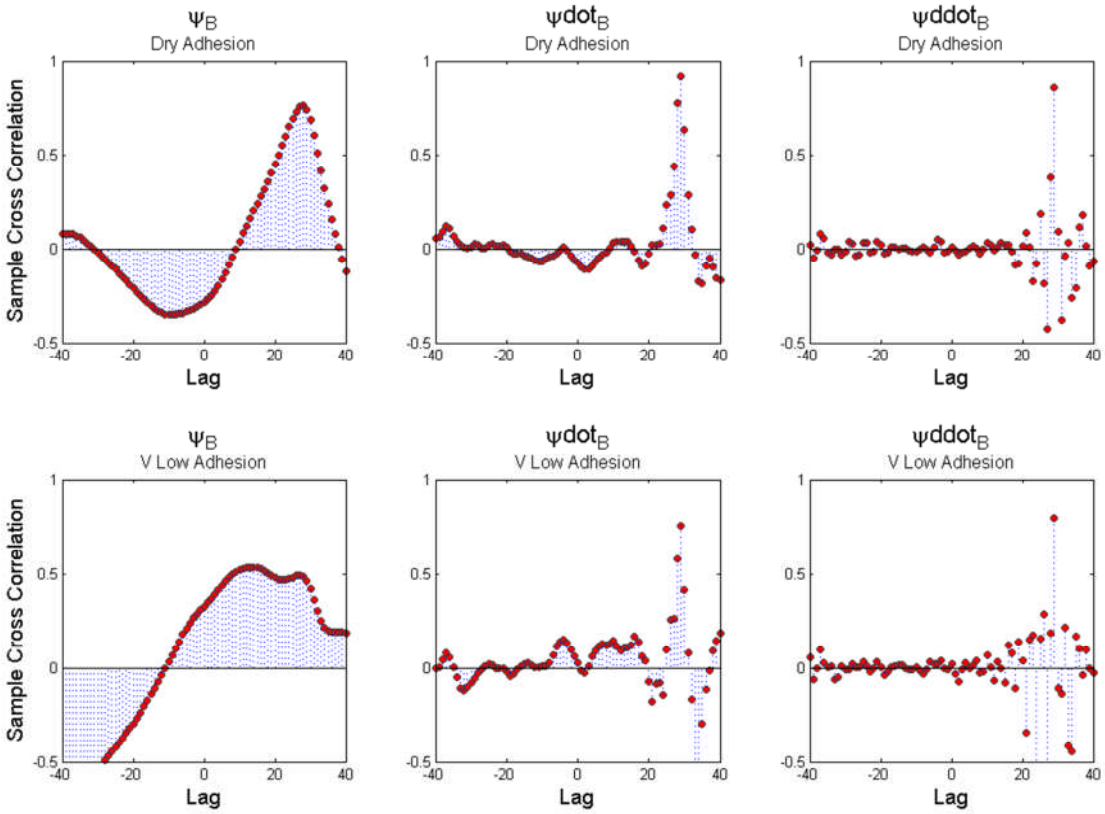


Figure 5. Correlation analysis between yaw dynamic for the leading and trailing bogie in both Dry (top row) and Very Low (bottom row) adhesion conditions

It can be seen that the yaw rate (ψdot_B) has a peak in correlation at a sample delay of 29 samples. This delay is concurrent with the time delay for the trailing bogie to reach the same place in the track as the leading bogie. It can also be seen that the peak is lower for the very low adhesion case than for the dry. It was proposed that a scheme using the correlation between leading and trailing bogie yaw rate (when the leading data is delayed by 29 samples) could be used to approximate adhesion.

3.2 Scheme for Adhesion Detection

Based on the conclusions from the correlation analysis, Figure 6 represents the proposed scheme for approximating adhesion. In this scheme the difference in value between leading and trailing bogie yaw rate (with appropriate delay included) has an RMS taken over a 10 second moving window. This value is then divided by the RMS of the leading bogie yaw rate to scale for the changing size in track irregularity.

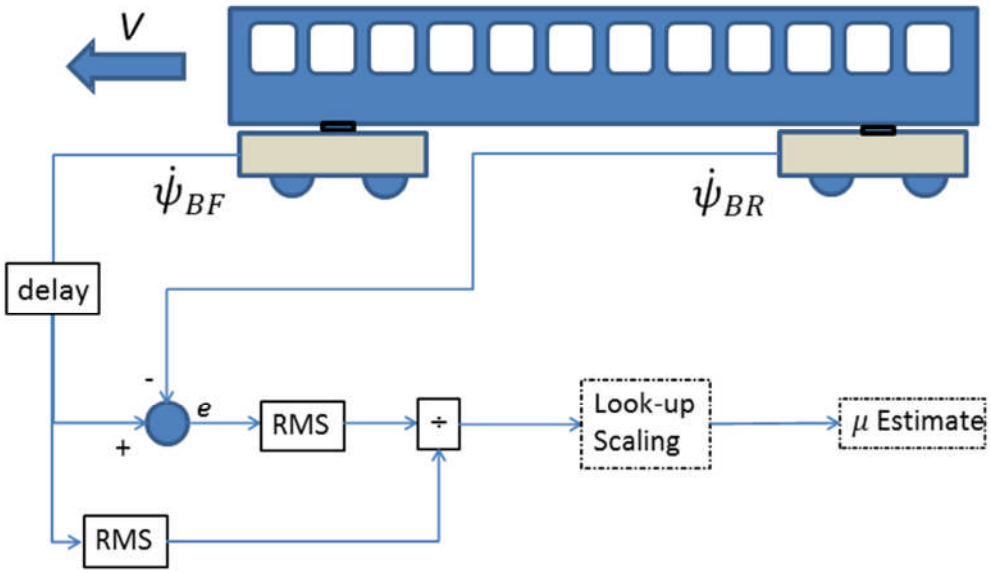


Figure 6. A direct data analysis approach for adhesion estimation.

This scheme was initially assessed over three test runs where the adhesion level was known. This allowed an average value to be taken and used to form a linear interpolation scheme based on the known adhesion levels.

3.3 Verification of Direct Data Approach Against VAMPIRE®

This method was tested against the same data as used in section 2.4, where a step change in adhesion level occurs 30 seconds into the test run. Figure 7 shows the result of this test.

It can be seen that this direct data method correctly identifies the step change and identifies the difference between the low and very low conditions. This method provides a slightly slower response as it was found a longer 10 second RMS window provided better results. However, these results have been obtained without the need to generate or verify a mathematical model.

As with the model-based approach, further verification of this method will again be performed using blind datasets.

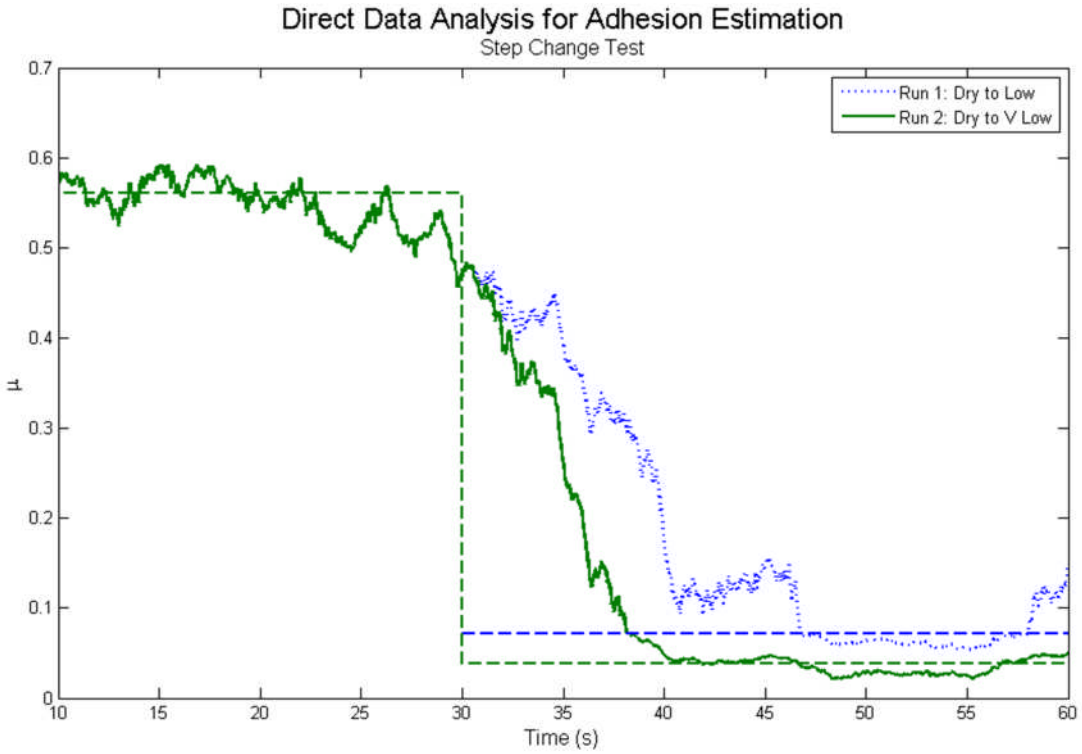


Figure 7. Figure showing direct-data based adhesion estimation for two step changes in adhesion.

5. Information Management

Both the model-based and direct data adhesion estimation methods provide reasonable results when approximating adhesion. However, there are two general features of each method that have to be taken into consideration. Firstly, both methods involve a time delay between passing over a section of track and analysing the adhesion level associated with it. Secondly, the accurate adhesion value is rarely correctly observed. Bearing this in mind, the usefulness of this data is assessed for both use on-board in real-time and as part of a rail network strategy.

5.1 Usefulness On-Board

The adhesion estimation ascertained was not found to be totally accurate with the known value. However it closely highlighted the level of operational risk associated with adhesion level. There is a distinct operational difference between low adhesion conditions and very low adhesion conditions as the former can be managed by defensive driving techniques. It is of operational usefulness to be able to highlight these different adhesion regions.

Because there is a delay between experiencing low adhesion and the quantification of it, there is a limited scope about using this measurement on-board. Conceptually, if an instantaneous reading could be found, it may be useful to inform the braking system controller to better manage the vehicle.

However, there has been some suggestion that small patches of low adhesion do not cause major operational issues, but longer, sustained patches do. Therefore it seems likely that a small reduction in analysis time to ~1-2 seconds would make a read-out to a train operator useful.

5.2 Rail Network Management

A more likely use of the captured adhesion level is the use of this system wide. If it were possible to map the experienced level of adhesion to geographical regions using a GPS system, it would be possible to identify high risk areas and target mitigation resource at these. A natural by-product of this method would be a historical map of problematic areas that would better inform the management of track conditions over seasonal changes.

However, this proposal is not without complication. GPS accuracy alone is such that it would be difficult to know which particular line is in use in a region where 2 or more tracks run side by side. Furthermore there is a conflict of interest between the cost of managing the on-board system and the benefit of the information. Currently, the cost of managing any additional computing on board will be absorbed by Train Operating Companies (TOCs) whereas the benefit will be realised by Network Rail.

6. Conclusion

This paper has shown two methods to approximate the adhesion of a rail track under normal operating conditions. The model-based approach carries a particular level of interest by the use of augmented states within a Kalman-Bucy filter. Although used here to approximate the size of creep force, it is possible that under the right circumstances, this method can be used to approximate values that cannot be directly measured in many other applications

The direct data approach used here has shown that a solution can be realised without the aid of a mathematical model, but by the observation of indirectly associated dynamics. However, it has also been shown that this method may not be transferrable to other bogie designs as a different scheme was needed for the older Mk3 bogie [6].

Both of these methods have produced reasonable adhesion estimations that observe the different operational risk levels under normal running conditions. The next stage of the project is to move away from simulation and verify the capability of these methods by a series of track tests.

Acknowledgements

This work was supported in part by RSSB under the project number T959.

Our acknowledgements go to TSLG for supporting this work, and RSSB for funding and managing the project and guiding it towards an appropriate solution to suit industrial needs. They also go to DeltaRail for providing

simulation test data and assisting in the generation of linear model for use in the model based estimator.

References

1. G. Charles, R. Goodall, and R. Dixon, “Model-based condition monitoring at the wheel–rail interface,” Vehicle System Dynamics, vol. 46, no. sup1, pp. 415-430, Sep. 2008.

2. C. Ward, R. Goodall, and R. Dixon, “Creep Force Estimation at the Wheel-Rail interface,” in Proceedings of the22nd International Symposium on Dynamics of Vehicles on Roads and Tracks, 2011.

3. C. P. Ward, R. M. Goodall, and R. Dixon, “Contact Force Estimation in the Railway Vehicle Wheel-Rail interface,” in Proceedings of the 18th IFAC World Congress, 2011.

4. D. I. Fletcher, “A new two-dimensional model of rolling-sliding contact creep curves for a range of lubrication types,” Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology, Nov. 2012.

5. J. J. Kalker, “A Fast Algorithm for the Simplified Theory of Rolling Contact,” Vehicle System Dynamics, vol. 11, no. 1, pp. 1-13, 1982.

6. P. Hubbard, “Real Time Detection of Low Adhesion in the Wheel / Rail Contact,” in RRUKA Annual Conference, 2012, no. November, pp. 1-5.

7. R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems 1,” vol. 82, no. Series D, pp. 35-45, 1960.

8. A H. Wickens, “The dynamics of railway vehicles—from Stephenson to Carter,” Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, vol. 212, no. 3, pp. 209-217, Jan. 1998.

9. V. K. Garg and R. V. Dukkipati, Dynamics of Railway Vehicle Systems. Academic Press, INC. (London) LTD., 1984, p. 483.

Use of Bayesian updating to combine experts' opinion and results of inspection in bridge management

Luis A.C. Neves
Nottingham Transportation Engineering Centre, University of Nottingham,
University Park, Nottingham, NG7 2RD, UK

Dan M. Frangopol
Fazlur R. Khan Endowed Chair of Structural Engineering and Architecture
Department of Civil and Environmental Engineering
ATLSS Center, Lehigh University, Bethlehem, Pennsylvania, U.S.A.

Abstract

Predicting future performance is fundamental in managing existing transport infra-structures, as bridges. Current practice uses the results of inspections as an indicator of current performance and predicts future deterioration based on Markov Processes. In this work, a probabilistic deterioration model proposed by Neves and Frangopol (2005) is used and combined with probabilistic inspection results. Bayesian updating is used to combine both sources of information, leading to a reduction in uncertainty in future performance predictions, without excessive belief in the results of inspections.

1. Introduction

Bridge management systems must combine the results of bridge performance assessment with performance deterioration models. In most cases, assessment must be carried out based only on visual inspections which, in spite of their low cost, provide very valuable information on current performance. These inspections can be associated with errors (Phares, Washer et al. 2004), resulting from the difficult condition under which inspections are carried out, inspectors' lack of experience, difficulties observing some structural elements, or human errors.

Deterioration of existing structures is a complex process, associated with very different mechanism and dependent on a manifold of environmental, material, geometry and use conditions. Consequently, significant uncertainty exists in future performance prediction. This uncertainty must be taken into account when making decisions in terms of maintenance and management, to avoid decisions increasing the long term risk.

The information gathered from visual inspections is complementary to that resulting from deterioration models. In fact, the first is accurate in defining present performance, as the latter is useful in predicting future performance evolution. Considering uncertainty exists in both sources of information, Bayesian updating can be used to define better predictions of performance, combining both data sets.

This methodology is applied to a set of reinforced concrete bridge elements, considering the deterioration model proposed by Neves and Frangopol (2005).

The obtained results show the impact of the use of Bayesian updating in reducing uncertainty, but avoiding excessive optimism regarding the quality of the inspection outcomes.

2. Bridge performance

Considering that visual inspections are the main source of information regarding the performance of existing bridges, most bridge management systems employ the results of such inspections as main indicator of performance (Thompson 1993, Hawk and Small 1998). Considering the qualitative nature of an inspection, most systems employ a discrete scale of values to define performance, and each component of a bridge is classified using a condition index or performance index.

The scales used to define the condition index vary enormously from component to component, and from country to country. However, in general, scales attributing the lower value to an intact structure and the higher value to a dangerous structure are used.

The deterioration of the condition index over time can be modelled using a manifold of approaches. The most common is based on the use of Markov models. This model considers a memory-less deterioration process, assuming a constant probability of transition between condition states. Neves and Frangopol (2005) proposed a continuous model for the condition index, assuming a bi-linear evolution curve. The deterioration of condition is defined in terms of three random variables: initial condition index, C_0, time to initiation of deterioration, t_{ic} and deterioration rate of condition, α_c, as follows:

$$C(t) = \begin{cases} C_0 & \text{if } t \le t_{ic} \\ C_0 + (t - t_{ic}) \times \alpha_c & \text{if } t > t_{ic} \end{cases} \quad (1)$$

The distributions used for each of these random variables can be defined based on statistical analysis of historical data or on experts' judgment.

Neves and Frangopol (2005) also employed a similar model for a safety indicator, considering that the random variables defining the safety index could be independent or correlated to the condition index. In this case, the safety index is defined as:

$$S(t) = \begin{cases} S_0 & \text{if } t \le t_i \\ S_0 + (t - t_i) \times \alpha & \text{if } t > t_i \end{cases} \quad (2)$$

where S_0 is the initial safety index, t_i is the initiation time of deterioration of safety, and α is the deterioration rate of safety. In the present work, the safety index is given by the bridge load capacity.

3. Bayesian updating

The models defined above are valid for a large set of structures or structural models similar to those used to calibrate the model. However, when a specific bridge is to be analysed, information on its present performance must be used. In existing bridge management systems, the result of the last inspection is considered accurate, and introduced as a deterministic value (Thompson 1993, Hawk and Small 1998). However, as discussed above, uncertainty exists on the result of an inspection and disregarding the possible variations from the expected result can lead to erroneous conclusions and decisions.

On the other hand, Bayesian updating can be used to combine the deterioration model with the results of inspections, assuming that a probability distribution can be defined for the condition index based on the result of an inspection.

Based on Bayes theorem, the probability density function of the condition index considering both sources of information can be defined as (Ang and Tang 2007):

$$f''(C_T) = K \cdot L(C_T) \cdot f'(C_T) \tag{3}$$

where $f''(C_T)$ is the probability density function of the condition index at time T considering both expert judgment and results of inspections, also designated posterior distribution, $f'(C_T)$ is the probability density function of the condition index at time T considering only expert judgment, also designated prior distribution, $L(C_T)$ is the likelihood function, and K is a normalizing constant defined by:

$$K = \frac{1}{\int_{-\infty}^{\infty} L(C_T) \cdot f'(C_T) dC_T} \tag{4}$$

The likelihood defines the probability of occurrence of a given condition index, C_T, knowing that an inspection result C_{ins} was obtained. Very little information exists on the uncertainty of inspection results, as this depends on the inspection procedures employed, the experience of inspectors, the characteristics of each bridge and component, and the inspection condition (weather, visibility, etc.). For this reason, the likelihood function is simplistically modelled using a Gaussian distribution, assuming no bias exists (e.g., the mean condition is assumed equal to the result of the inspection). The standard deviation of the likelihood is defined considering levels of inspection quality (Neves and Frangopol 2008, Neves and Frangopol 2010).

Since the condition index probabilistic indicators are computed using Monte-Carlo simulation, the same method is applied for Bayesian updating.

Considering that all samples have a probability of occurrence equal to 1/*n*, where *n* is the number of samples, equation (3) becomes:

$$f''(C_T) = K \cdot L(C_T) \cdot \frac{1}{n} \tag{5}$$

and expression (4) becomes:

$$K = \frac{1}{\sum_{i=1}^{n} L(C_T^i) \frac{1}{n}} \tag{6}$$

The mean, square mean and standard deviation of the updated condition index can be computed using a weighted average procedure as (Chen and Ibrahim 2000):

$$\mu_C^\tau = \sum_{i=1}^{n} C_\tau^i \cdot P''(C_T^i) = \sum_{i=1}^{n} C_\tau^i \cdot K \cdot L(C_T^i) \cdot \frac{1}{n} = \frac{\sum_{i=1}^{n} C_\tau^i \cdot L(C_T^i)}{\sum_{i=1}^{n} L(C_T^i)} \tag{7}$$

$$\mu_{C^2}^\tau = \sum_{i=1}^{n} (C_\tau^i)^2 \cdot P''(C_T^i) = \sum_{i=1}^{n} (C_\tau^i)^2 \cdot K \cdot L(C_T^i) \cdot \frac{1}{n} = \frac{\sum_{i=1}^{n} (C_\tau^i)^2 \cdot L(C_T^i)}{\sum_{i=1}^{n} L(C_T^i)} \tag{8}$$

$$\sigma_C^\tau = \sqrt{\mu_{C^2}^\tau - (\mu_C^\tau)^2} \tag{9}$$

If, following Neves and Frangopol (2005), the safety index profile is assessed in parallel with the condition index, the updated safety index can be computed using the above equations, replacing C_τ^i by S_τ^i.

3. Applications

The proposed method is applied to a set of reinforced concrete elements in the United Kingdom. The probabilistic distributions of parameters defining the condition index and safety index are based on experts' judgment (Denton 2002), as presented in Table 1. All probabilistic parameters defining the condition index and safety index are described by triangular distributions, in terms of the corresponding minimum, mode and maximum values.

Considering no maintenance or inspection, the predicted condition and safety indices are as presented in Figures 1 and 2, respectively. These figures show that the standard deviation of both the condition index and the safety index increase rapidly, and a long term prediction of performance is associated with very large uncertainty.

	Initial Index	Time of Damage Initiation (years)	Deterioration Rate (year^{-1})
Condition Index	0.00 1.75 3.50	0	0.00 0.08 0.16
Safety Index	0.91 1.50 2.5	0	0.00 0.015 0.035

Table 1 – Condition index and Safety index parameters under no maintenance (Denton 2002)

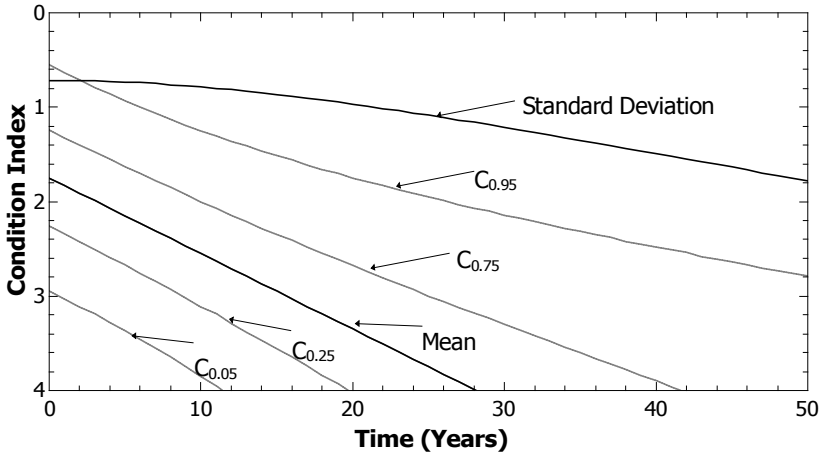


Figure 1. Mean, standard deviation and percentiles of the condition index not considering inspections (Neves and Frangopol 2005)

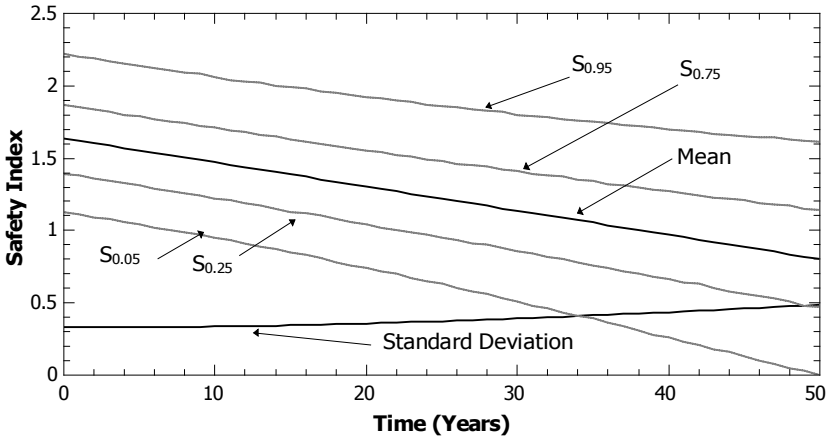


Figure 2. Mean, standard deviation and percentiles of the safety index not considering inspections (Neves and Frangopol 2005)

3.1 Condition Index updating

As a first application, it is assumed that an inspection is carried out at year 10. It is considered that the outcome of the inspection, C_{ins}^{T}, is either 2 or 3. In terms of standard deviation, a set of inspection quality levels, as well as, the corresponding likelihood function and probability of misclassification are presented in Table 1.

Quality	Probability of misclassification	Mean Value	Standard deviation
High	5%	C_{ins}	0.255
Medium	10%	C_{ins}	0.304
Very Low	40%	C_{ins}	0.595

Table 1 – Probability distribution of the likelihhod function in terms of the quality of inspection (Neves and Frangopol 2008).

The mean and standard deviation of the updated condition index considering different inspection quality levels and an inspection result equal to 2 and 3 are presented in Figures 3 and 4, respectively.

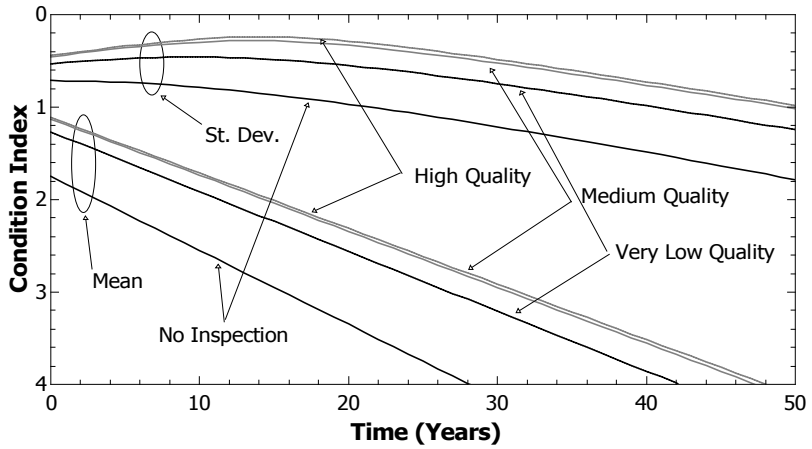


Figure 3. Mean and standard deviation of the condition index not considering inspections and different inspection quality levels with $C_{ins}^{16} = 2$

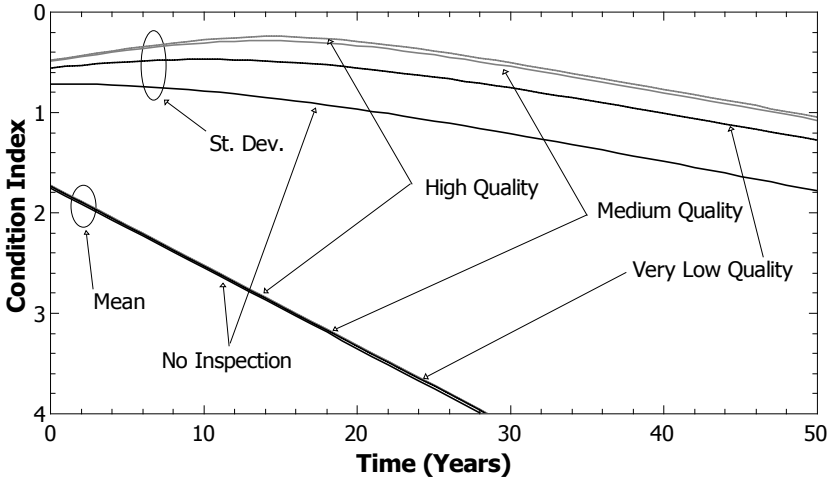


Figure 4. Mean and standard deviation of the condition index not considering inspections and different inspection quality levels with $C_{ins}^{16} = 3$

As shown in Figure 3, the result of the inspection, $C_{ins} = 2$, is better than initially predicted. As a result, the updated mean condition is lower (i.e., less deteriorated) than the initial prediction. This improvement increases with the quality of the inspection, since growing confidence in the results of the inspection result in higher weight of this result.

In terms of standard deviation, a significant reduction is observed for all inspection quality levels. The minimum standard deviation is observed close to the time of inspection, and the uncertainty in the prediction grows with the elapsed time since inspection.

In Figure 4, the results obtained considering an inspection with an outcome, $C_{ins} = 3$, are presented. In this case, the outcome of the inspection is very similar to the initial prediction, resulting in a significantly smaller impact of the updating procedure in the mean condition index. However, in terms of the standard deviation, a similar reduction to that observed in the previous example is present.

A significant advantage of this approach in comparison to common practice, where the result of a single inspection is assumed as the true condition index of the structure at time of inspection, is the ability to consider several inspections at different instants.

Considering two high quality inspections are carried out at year 16 and 36, with results $C_{ins}^{16} = 2$ and $C_{ins}^{36} = 3$ results in the updated condition index shown in Figure 5. Comparing these results with those obtained considering no inspections and one high quality inspection, shows a significant change in the expected deterioration rate, as well as, a reduction in standard deviation of the condition index.

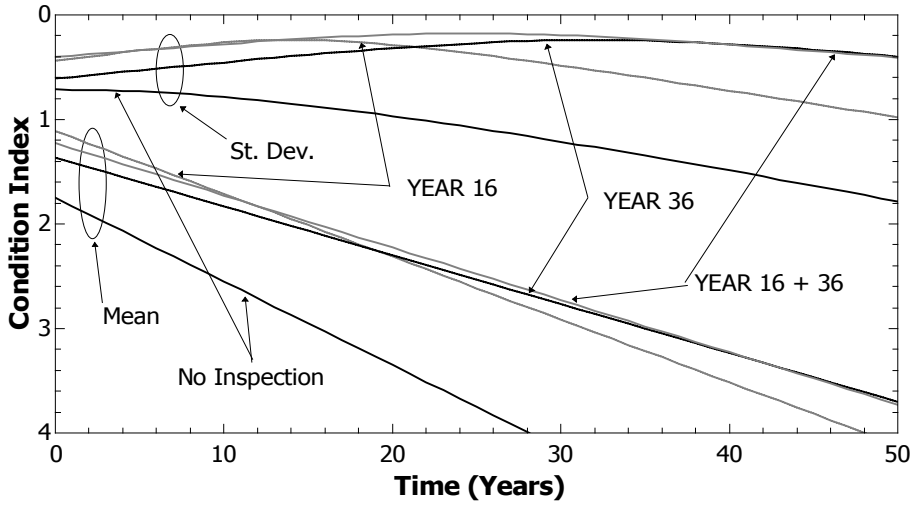


Figure 5. Mean and standard deviation of the condition index not considering inspections and one or two inspections at years 16 and 36

In fact, at year 16, the uncertainty in the condition index is mostly associated with the initial condition, C_0, and an inspection has little influence on the prediction of the deterioration rate, α_c. However, as deterioration progresses,

the influence of the deterioration rate increases, and a later inspection will have a much greater influence on this parameter. Moreover, the use of two inspections provides a greater insight regarding the deterioration process between the times of inspection. These results are illustrated in Figures 6 and 7, where the initial condition index and deterioration rate of condition distributions are presented considering: (i) no inspection; (ii) an inspection at year 16; (iii) an inspection at year 36; and (iv) two inspections at year 16 and 36. In all cases, a high quality inspection was considered and the results of inspections are taken as $C_{ins}^{16} = 2$ and $C_{ins}^{36} = 3$.

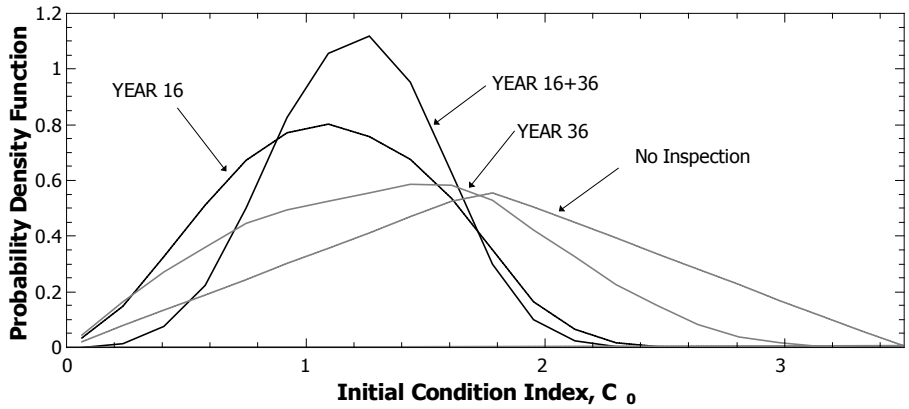


Figure 6. Distribution of the initial condition index not considering inspections and one or two inspections at years 16 and 36

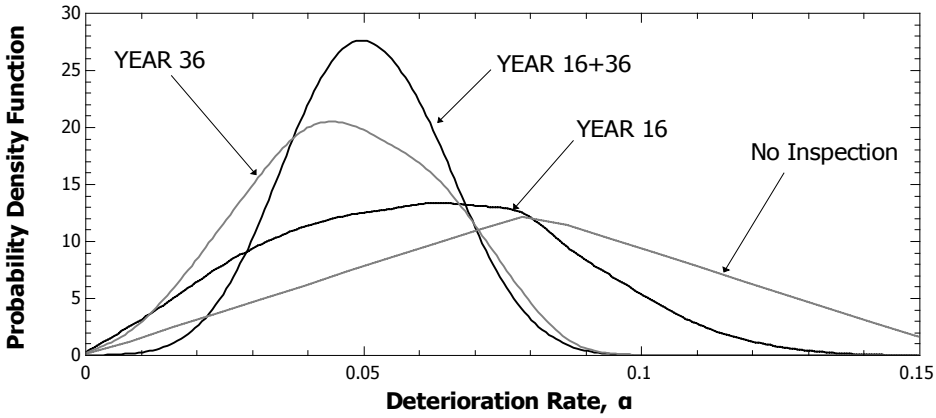


Figure 7. Distribution of the condition deterioration rate not considering inspections and one or two inspections at years 16 and 36

3.2 Safety Index updating

Although inspections produce no direct information regarding the safety of a bridge or component, it is reasonable to consider that the deterioration of condition and safety are correlated processes. In this case, information on the condition index of a bridge or component will provide indirect information regarding its safety.

As an example, the same data the same inspection procedures defined above, considering one or two high quality inspections will be employed. Assuming a correlation between the initial condition index and initial safety index, as well as, between the deterioration rate of the condition index and the safety index equal to 0.4, the updated safety index is as shown in Figure 8.

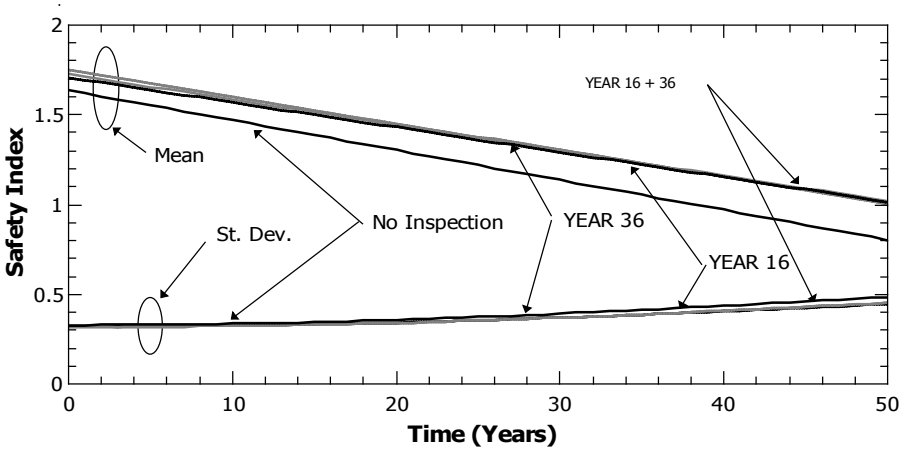


Figure 8. Mean and standard deviation of the safety index not considering inspections and one or two inspections at years 16 and 36

These results show that, since the observed condition index is better than the initial prediction, an improvement in the safety index also results. In terms of standard deviation, a reduction is observed, although much less significant than that observed in the condition index.

4. Conclusions

In this work, a methodology for combining the results of bridge inspections with deterioration models is presented. Both sources of information are probabilistic and a Bayesian updating methodology, using Monte-Carlo simulation, is employed.

The proposed model is applied to a set of reinforced concrete bridge elements, characterized using experts judgment (Denton 2002). The performance of these elements is analysed considering two indicators: condition index, resulting directly from visual inspections, and safety index, resulting from structural and safety analysis.

Performance is evaluated not considering inspections, and compared to those assuming different inspections scenarios, in terms of number and outcome of inspection and quality of the inspection procedure.

Results show that a significant reduction in performance uncertainty occurs when inspections are considered, in particular for the condition index. Although this improvement is greater for high quality inspections, even when significant probabilities of misclassification are considered, an important improvement is observed.

References

1. Ang, A. H. S. and W. H. Tang. Probability Concepts in Engineering: Emphasis on Applications in Civil & Environmental Engineering, Wiley (2007).
2. Chen, M. H. and J. G. Ibrahim. Monte Carlo Methods in Bayesian Computation, Springer (2000).
3. Denton, S.. Data Estimates for Different Maintenance Options for Reinforced Concrete Cross-Heads, Brinckerhoff Ltd (2002).
4. Hawk, H. and E. P. Small. The BRIDGIT bridge management system. *Structural Engineering International*, IABSE **8**(4): 303-314 (1998).
5. Neves, L. C. and D. M. Frangopol. Condition, safety and cost profiles for deteriorating structures with emphasis on bridges. *Reliability Engineering & System Safety* **89**(2): 185-198 (2005).
6. Neves, L. C. and D. M. Frangopol. Life-cycle performance of structures: combining experts judgment and results of inspection. *Life-Cycle in Civil Engineering*, Varenna, Italy (2008).
7. Neves, L. C. and D. M. Frangopol. Optimization of bridge maintenance actions considering combination of sources of information: Inspections and expert judgment. *The Fifth International Conference on Bridge Maintenance, Safety and Management* (2010).
8. Phares, B. M., G. A. Washer, D. D. Rolander, B. A. Graybeal and M. Moore. Routine Highway Bridge Inspection Condition Documentation Accuracy and Reliability. *Journal of Bridge Engineering* **9**(4): 403-413 (2004).
9. Thompson, P. D.. The Pontis Bridge Management System. Pacific Rim TransTech Conference: International Ties, Management Systems, Propulsion Technology, Strategic Highway Research Program, Seattle, ASCE (1993).

Stochastic State Space Methods for Railway Network Asset Management Modelling

Darren Prescott, John Andrews
Nottingham Transportation Engineering Centre,
University of Nottingham, NG7 2RD, UK

Abstract

If the UK population increases in line with predictions over the coming decades there will be an associated increase in the requirement for road and railway networks to deal with increased levels of passenger and freight transport. Greater utilisation of the railway network will bring two major, related problems: firstly, the increased level of traffic will lead to a faster deterioration in track quality; secondly, the available windows for maintenance will become shorter and less frequent. Even without these two problems, there is already great pressure to maximise the cost-effectiveness of the railway whilst ensuring that services remain efficient and free from disruption.

Asset management is therefore of vital importance to the railway network; activities such as maintenance and renewal must be carefully planned to ensure an acceptable level of service performance. Accurate asset management models are vital in developing strategies to support effective asset management decision making. This paper outlines and compares two asset management models that have been developed to model the deterioration, inspection and maintenance of railway track and considers their application to models of the wider UK rail network.

1. Introduction

Railway track geometry deterioration occurs due to the cumulative effect of traffic on the railway and has a direct influence on passenger comfort. Without intervention it is possible for track geometry to degrade to such an extent that there are also serious safety implications due to the increase in the risk of derailment. Therefore, if the track geometry falls significantly action must be taken to restore the track to an acceptable condition. Standards specify that the higher the speed limit of the track, the lower the acceptable level of track geometry degradation. Trains equipped with special measuring equipment are used at regular intervals to determine the track geometry and hence assess the level of deterioration. If the track quality falls below the acceptable levels its geometry is restored by adjusting the track ballast, through either manual intervention, tamping or stoneblowing. It is important to have a thorough understanding of the track deterioration and maintenance processes so that sound decisions can be made as to how and when to maintain ballast and hence to ensure safe, efficient and effective operation of the railway.

A number of models have been developed for use as support tools in track asset management. These models aim to predict the track geometry condition and can be classified as being either deterministic or stochastic.

Deterministic models were developed by both Shenton [1] and Chrismer and Selig [2] to describe the settlement of ballast, which they considered to be the main factor in track geometry degradation. Sato [3] derived a deterministic model to express track deterioration in terms of the average growth of track irregularity and Hamid and Gross [4] expressed the state of the track using artificial track quality indices (TQI). These indices give a simple indication of the track state but have no physical meaning and are thus of questionable use in the context of track asset management. TQI's have also been proposed by Bing and Gross [5] and Askarinejad [6].

Stochastic models of the changing track geometry condition include the Markov model developed by Shafahi et al [7], which uses a TQI on a scale of 0-100 based on measurements of the track unevenness, twist, alignment and gauge. The TQI was mapped on to 5 states in the Markov model and the elements of the transition probability matrix were then calculated from the variation of the TQI with time. Lyngby et al [8] proposed a 50-state Markov model representing the change in track twist over time, where each state represents the twist of a piece of track up to 50mm in length. Deterioration rates were specified according to whether the track was straight, curved or a transition between the two. The model was used to optimise the frequency of track geometry inspections. Podofillini et al [9] and Kumar et al [10] also applied RAMS approaches to rail failure modelling.

In order to assist in track asset management decision making the possible track maintenance and renewal processes must be integrated with the track degradation process to produce a track asset management model. This model can then be used to investigate the effects of different maintenance and inspection regimes and evaluate their effectiveness. A limited number of models have been developed to evaluate the effects of maintenance on track geometry. Recent work by Quiroga and Schnieder [11,12] used data from the French railway operator, SNCF, to produce statistical models of the form:

$$Q = Ae^{B(t-t_0)} + \varepsilon(t) \tag{1}$$

where Q is the track quality measure and A, B, and ε are parameters assumed to have lognormal, normal and normal distributions respectively and t_0 is the time of the last intervention (tamping) activity. After establishing the model parameter distributions Monte Carlo simulation was used to evaluate the model and establish an optimal level of performance.

Modelling approaches such as Petri nets and Markov analysis are able to explicitly take account of many track maintenance and renewal options. As such, they are extremely well suited to the investigation of track asset management. Petri net approaches such as those employed by Andrews [13] and Prescott and Andrews [14] can be used in conjunction with Monte Carlo simulation to analyse scenarios where the track geometry deterioration can be shown to vary with time. Markov analysis has also been used by Prescott and Andrews [15] to model the lifetime of a track section subjected to tamping at maintenance interventions. In this paper the Petri net and Markov approaches are compared and their relative benefits for the support of track asset

management are discussed, along with their potential to be used as a basis for whole-system asset management models of the entire railway network.

2. Track Geometry

Track geometry is influenced by the weight, speed and number of trains that use the railway. Instrumented measurement trains are used at regular intervals to monitor the geometry of the rails and the recorded data is then used to produce indications of the track geometry over short track sections; in the UK these track sections are one eighth of a mile in length. The data is processed to derive the following measurements:

- Vertical geometry (top) - rail height standard deviation (SD) (left, right and mean), subtracted from a 35 m (short-wave) or 70 m (long-wave) running average.
- Lateral geometry (alignment) - rail alignment, averaged over the left and right rail, expressed as short and long-wave SDs.
- Gauge - rail spacing compared to the standard gauge.
- Cyclic top - a measure of resonant dip frequencies (which may cause a train to “bump” along the track, possibly leading to a derailment).
- Twist - a measure of cant variation over 3 m and 5 m.

Acceptable limits are specified for each of these measurements according to the speed of the track but the short-wave measurement of vertical geometry is considered to be the main indicator of track quality.

3. Track Maintenance

Any track section whose geometry is found to lie outside the acceptable limits must be maintained. This maintenance restores the track alignment by adjusting the stone ballast under the sleepers so that the rails are parallel and at a relative level that is appropriate to the track curvature. The maintenance can be performed manually in urgent situations over short lengths of track but in all other cases machines (either tampers or stoneblowers) are used.

Tamping machines work by measuring the existing track geometry, lifting the track precisely up to the required level and then inserting large, vibrating tines into the ballast either side of a sleeper. The tines then squeeze the ballast, packing it tightly below the sleeper, so that when the machine releases the track it is at the required level. However, in addition to improving the track alignment, tamping also results in degradation of the ballast condition, since the action of the tines in the ballast breaks up some of its constituent stones, leading it to perform less effectively following each tamping intervention.

Stoneblowers work in a similar way to tampers, in that they measure the track geometry and lift the track to the required level. However, rather than disturbing the existing ballast, stoneblowing involves the pneumatic injection of small stones under the sleeper to hold the track in the require position. Once stoneblowing has been performed tamping must not be carried out since it would result in the small stones working their way down through the ballast, affecting its mechanical properties and its drainage.

Figure 1 shows a typical plot of the vertical alignment readings taken by a measurement train on a track section, along with the recorded times of renewal and three successive tamping interventions. The deterioration of the track quality can be inferred from the increasing value of the reading with time and the effect of the tamping interventions in improving the track quality as determined by this measure is also clearly visible.

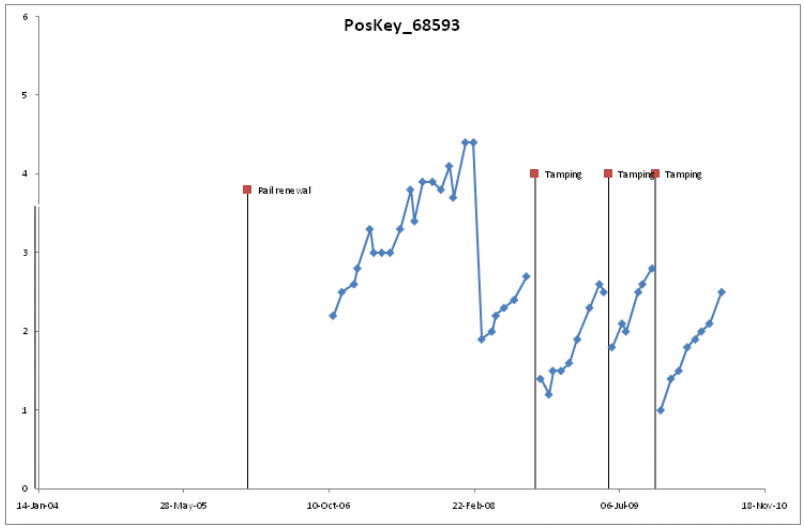


Figure 1. Example plot showing measured values of the maximum averaged short-wave average top standard deviation, renewal and tamping interventions for a 1/8th mile track section.

4. Analysing Deterioration Trends

A general model representing the condition of a track section is shown in Figure 2, which shows how the standard deviation (SD) of a track quality measurement changes with time. The track does not return to an as good as new (AGAN) condition after tamping (T) and the rate of deterioration changes following each tamp. σ_{Crit} is a critical value of the SD, which leads to a request for maintenance to be placed for this section. The measurement train is used every θ days and it is not until the track condition is measured at time t_{in_req} that it is discovered that maintenance is necessary. After a time delay, D, the tamper carries out maintenance.

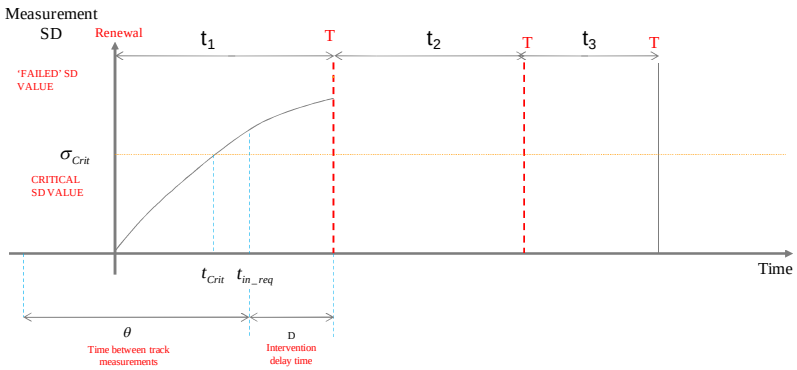


Figure 2. A track section condition model.

The track deterioration process is dependent on the maintenance history, meaning that analysis of the deterioration must be considered in separate phases according to the number of interventions that have taken place, as shown in Figure 3. The statistical analysis used to determine the times to reach any specified level of deterioration in any life phase is presented in reference 13. The deterioration time distributions determined using this statistical analysis are used in the track asset management models.

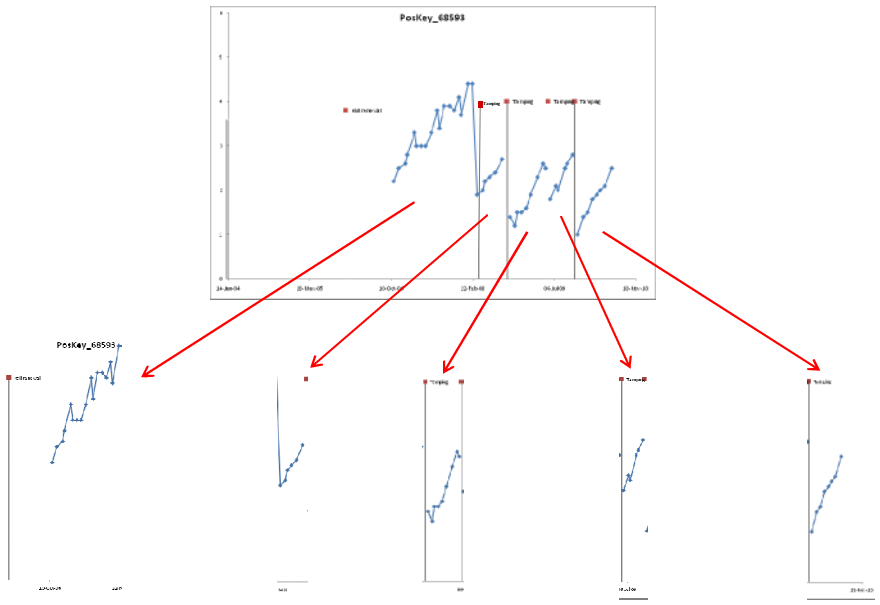


Figure 3. Identifying the phase of deterioration.

5. Track Asset Management Models

In this section two track asset management models, one based on Markov Analysis and one based on Petri nets, are described following an initial description of the assumptions they share.

5.1 Common Assumptions

The models presented in this section make a number of common assumptions:

- The track is classified according to the measured SD of the vertical alignment, σ, as follows:
  $0 \leq \sigma < \sigma_{crit}$ good condition,
  $\sigma_{crit} \leq \sigma < \sigma_{spd}$ maintenance requested,
  $\sigma_{spd} \leq \sigma < \sigma_{cls}$ speed restriction required,
  $\sigma_{cls} \leq \sigma$ line closure required.
  The values σ_{spd} and σ_{cls} are set according to the speed band of the track being analysed and the value of σ_{crit} can be varied.
- A track measurement train (TMT) measures the track geometry, and hence reveals its present condition, at set intervals of θ days.

- Maintenance can be either of a normal or an urgent priority; urgent priority maintenance is needed if line closures or speed restrictions must be imposed.
- The track's deterioration process changes during the different phases of its lifetime, as shown in Figure 2.

5.2 Markov Track Model

Markov Analysis: Markov models are state-space models containing nodes, which represent distinct states of the system being modelled, and directed edges between the nodes, which indicate how the modelled system moves between states. In the analysis presented the models are Markov processes and hence each edge has associated with it a value representing the rate at which the system will move from the start node to the end node. An example Markov model is shown in Figure 4 for a component that can be either working (W) or failed (F), has failure rate λ and repair rate ν. Markov models such as this result in the production of a system of differential equations that can be solved to give either the steady state or transient probability of being in each of the model states. Methods of analysis can be found, for example, in reference 16.

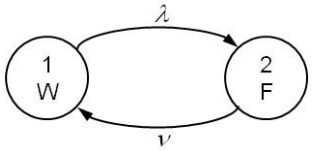


Figure 4. State space diagram for a Markov model of a two state-repairable component.

Model Assumptions: A Markov model for a single section of railway track was first presented in reference 15. In addition to the assumptions outlined in Section 5.1 this model also involves the following assumptions:

- Constant rates of deterioration of the track from one state (of deterioration) to another, e.g. good condition to a condition where maintenance is requested, during any particular life phase.
- Maintenance involves tamping only. After 7 tamps the rates of deterioration between states no longer changes in the way illustrated in Figure 2 but instead stays constant.

The first of these assumptions is necessary due to the fact that Markov analysis is being applied; to use the Markov approach the system must be stationary (in addition to being memoryless). In this case, the second assumption was made because of the track data that was available.

The Track Section Model: Figure 5 shows the general form of the Markov model that represents the deterioration and inspection of the track section following renewal. During deterioration the actual (A) condition of the track worsens (solid edges) and during inspection the known (K) condition of the track is revealed (dashed edges). The deterioration rates between the different track states are λ_{ij}, where *i* represents the deterioration level (1: good

to crit, 2: crit to spd and 3: spd to cls) and j represents the number of tamps that have been performed on the track section since the ballast was renewed. The λ_{ij} are calculated by taking the reciprocals of the average times taken for the track section to deteriorate from one condition to the next worst after a specific number of tamps. These average times were derived from real network data.

Since the deterioration process changes following each maintenance intervention, the track section maintenance cannot take the track back to an AGAN condition as represented by state 0 in Figure 5. Therefore, the Markov model for the track section must contain a group of model states of the form shown in Figure 5 for each phase of the section's lifetime (after renewal, 1 tamp, 2 tamps and so on). Since there are 8 life phases (after renewal and after each of 7 tamps) there are 80 model states, which are numbered from 0 to 79 using the formula $10j + k$, where j is the number of tamps performed (if the number of tamps ≥ 7 then $j = 7$) and k is the track condition as represented in the model states shown in Figure 5. The degradation and inspection transitions between the groups of 10 model states representing the different life phases are similar to those shown in Figure 5. The difference lies in the deterioration rates, λ_{ij}, which vary according to the life phase of the section, with j determined according to the number of tamps that has been performed.

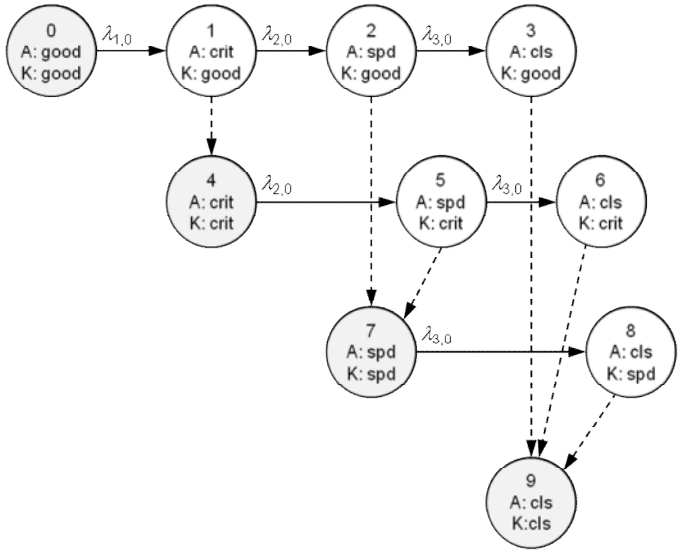


Figure 5. Markov model for track section degradation (solid arrows) and inspection (dashed arrows) following track renewal; shaded model states show the revealed track conditions.

Figure 6 shows the model states that represent the first two phases of the track section's lifetime: after renewal (states 0-9) and after a single tamp (states 10-19). Edges representing maintenance are included on this diagram, with rate ν_{norm} for normal priority maintenance and ν_{urg} for urgent priority maintenance. These lead from states where maintenance is requested (states 4-9 and 14-19) to the state in the next phase of the track section's life where the track is in a good condition (state 10 after 1 tamp and state 20, not shown,

after 2 tamps). Edges representing normal priority maintenance lead from each of the states where the known (K) condition of the track has exceeded σ_{crit}, the level of SD at which maintenance is requested, to the state representing a good condition in the next phase of the section's life. The actual (A) condition of the track could be worse but the maintenance will not be of an urgent priority until that condition is revealed at inspection. Edges representing urgent priority maintenance lead from each of the states where the known condition of the track requires either speed restrictions or line closures to the state representing a good condition in the next phase of the section's life. Since it is assumed that maintenance after 7 tamps or more returns the track to the same condition (i.e. the deterioration rates between states no longer changes) the edges representing maintenance from model states 74-79 lead back to state 70, which represents the track being in a good condition after 7 tamps.

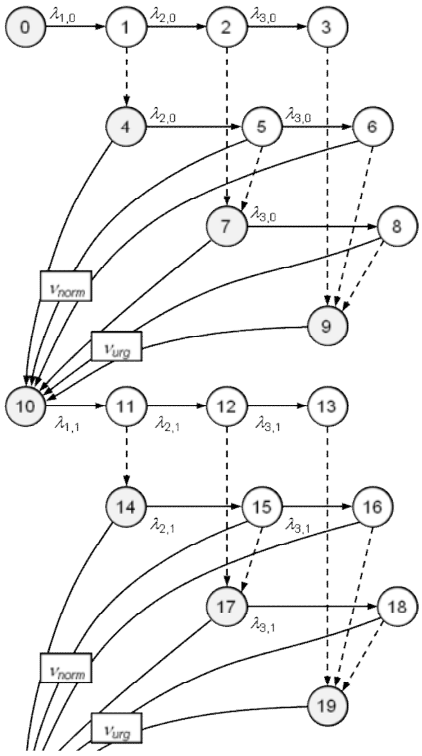


Figure 6. The part of the track section Markov model showing deterioration, inspection and maintenance after renewal and after a single tamp.

Model Solution: The Markov model contains 80 states, which lead to a system of 80 differential equations. This system of equations can be solved numerically using a solution routine such as Runge-Kutta in order to obtain the transient system behaviour, i.e. the probabilities of being in each of the 80 system states throughout the integrated time period. Each of the deterioration and maintenance transitions have associated rates, as previously described, and the inspection transitions can be modelled by shifting probabilities between states at the appropriate times in the integration process. Therefore, at intervals of θ days during the integration, state probabilities are adjusted to account for the inspection. For example, referring to Figure 5, when an

inspection takes place the probability of being in state 4 will rise by the probability of being in state 1, which will itself then be reduced to zero; this signifies that the state of the track is revealed by the inspection train, so that the actual and known conditions match.

5.3 Petri Nets

The Petri Net Method: A Petri net (PN) is a directed graph consisting of two types of nodes: places and transitions, linked by directed edges from either place to transition or transition to place. An example PN is shown in Figure 7, the circular nodes represent places and the rectangular node represents a transition. In reliability analysis the places in a PN represent the states of a system and the present state is known according to the PN marking, the number of tokens contained by the various PN places. The marking of the PN changes (and, hence, so does the system state) when tokens are redistributed throughout the PN as transitions are enabled and then fire. A transition is enabled when all of its input places are marked; once enabled it will fire after the delay time associated with it has elapsed (the transition in Figure 7 has delay t) and a token will be removed from each of the input places and one added to each of the output places.

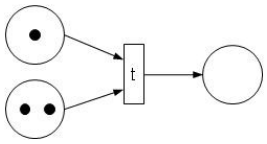


Figure 7. An example Petri net.

Model Assumptions: A PN model that could be applied to multiple track sections within a railway network was presented in reference 14 (a model for a single track section was presented in reference 13). In addition to the assumptions outlined in Section 5.1 this model also involves the following assumptions:

- The rail network is split into a number of regions, each of which contains a number of track sections.
- The maintenance engineer in each region decides on the maintenance to be performed in that region given the condition of the track sections within it.
- Maintenance is carried out by 2 tampers and 1 stoneblower, which are shared across the entire network.
- Tamping is preferable to stoneblowing, since once a section has been stoneblown the tamper can no longer be used. However, urgent maintenance is performed by whatever machine is first available and if a tamper has not become available after a specified time, the stoneblower will be used, if available.
- An extra threshold, σ_{opp}, of the measured SD of the vertical alignment is introduced ($\sigma_{opp} < \sigma_{crit}$). If the measured SD is above this opportunistic maintenance can be performed if a maintenance machine is in the area.

It can be seen from the extra assumptions outlined above that the PN model aims to analyse a far more complex system than the Markov model. As would be expected, the PN model is not bound by the limiting, stationary assumption of the Markov model, meaning that non-exponential deterioration distributions can be considered. From analysis of actual track data, it seems that these distributions are a better match for reality.
The PN Model: The PN model developed in reference 14 contains a number of separate modules, which relate to the various track sections and regions. Section modules cover the degradation, inspection, maintenance and the identification of both the urgency of maintenance and which machines can possibly perform it. Regional modules keep track of the number of sections to be maintained and how and whether maintenance machines are available for use in the region; these modules also take account of the decision making made by the regional maintenance engineer. Here, PN modules for the deterioration and inspection of the track section are presented, in order to provide a comparison with the presented Markov model.

Figure 8 shows the PN module for track section deterioration. As marked, the track is in a good state, with transition T1 enabled. When T1 fires after time delay t_1, sampled from the appropriate deterioration distribution, P2 will be marked, signifying that the track has deteriorated to the level where opportunistic maintenance might now be performed. Places P6 and P7 count the number and type of previous maintenance interventions and this maintenance history affects the distributions from which delay times t_1-t_4 are sampled in the place conditional transitions[17] T1-T4, thus accounting for the change in deterioration time distribution as illustrated in Figure 2.

Figure 9 shows the module for track section inspection. Here, the TMT inspection process is modelled to the right of the PN (P101-T6-P102-T5), with the actual state of the track condition (P2-P5) revealed if necessary at inspection (P101 marked) with P6-P11 signifying that the track condition is known.

The modules shown in Figures 8 and 9 cover the entire deterioration and inspection process modelled in the 80 states of the Markov model (which related to the 8 life phases of renewal and the following successive tamps up to 7). However, the PN modules shown model more than this; they can account for any number of both tamps and stoneblows following renewal, with no increase in the model size.

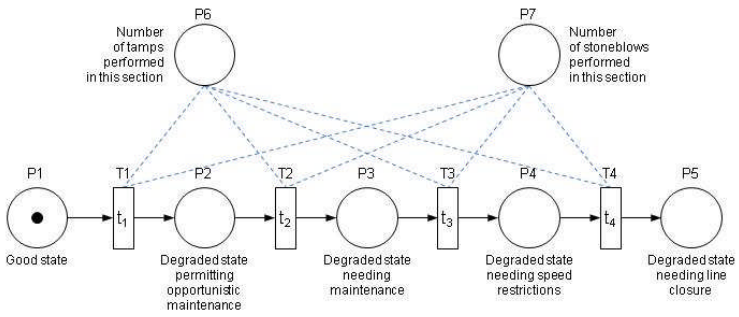


Figure 8. PN module for track section deterioration.

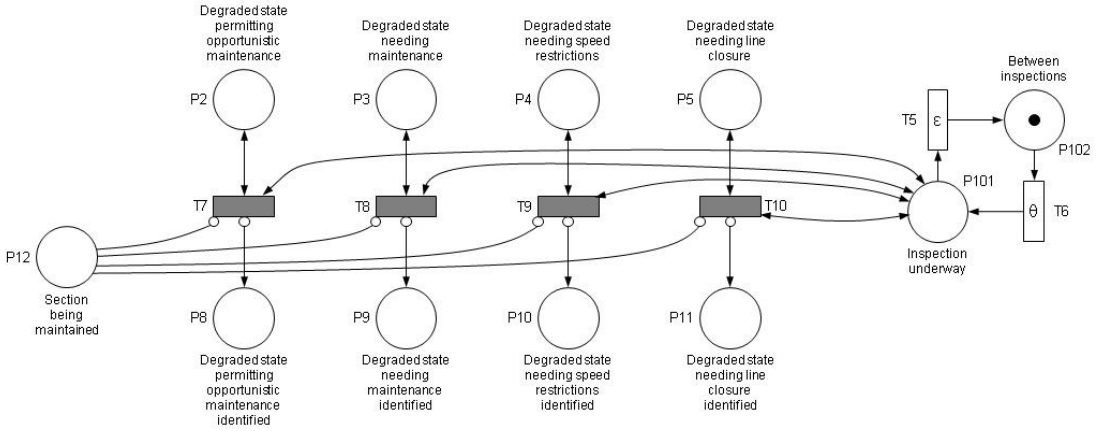


Figure 9. PN module for track section inspection.

Model Solution: The completed PN model forms an ideal framework for Monte Carlo simulation. Decision making at maintenance involves marking appropriate places in the PN when a maintenance event is simulated in order to signify that specific sections are to be maintained by specific machines. The times at which the track sections move from one state to another are determined by sampling the relevant deterioration time distributions according to the track section maintenance history. It is possible to collect a variety of different information from the PN model, by monitoring the number of times that specific places are marked and for how long. For example, in addition to monitoring the probability of the track being in different states, it would be possible to investigate other relevant metrics such as maintenance machine utilisation and associated costs, type and distribution of maintenance delays, the length of time different track conditions are undetected and so on.

6. Discussion

Two railway track asset management models have been presented and their major features outlined. Each has its advantages.

The Markov model is a simple model, which can be quickly evaluated to give a reasonable indication of the effects of a chosen maintenance policy. However, in common with all Markov models, the main problem with the implementation of the presented model is the potential state space explosion. The presented model for a single track section only considers a single maintenance machine but already contains 80 states. If multiple track sections and further maintenance machines were to be considered, then the model would soon grow impossibly large to analyse.

The discussed PN track model is capable of modelling a more complex situation than that which can be modelled using the Markov model. A good illustration of this is provided by the fact that the two PN modules presented in Figures 8 and 9 model all details covered by the 80-state Markov model. In addition to this, despite the fact that the PN model is considerably more compact, it also models a level of SD at which opportunistic maintenance is

triggered and a maintenance history accounting for the actions of a number of tamping and stoneblowing machines (note that, in contrast to the Markov model the size of the PN model does not change no matter what type and how many maintenance interventions are considered). The modular nature of the PN is also of great benefit in scaling up the model to cover the railway network. Modules are easily replicated for new track sections, thus allowing the model to be relatively simply expanded to a larger size. This is in stark contrast to the Markov model. Perhaps the major disadvantage of the PN model is the fact that it is solved using a Monte Carlo simulation solution routine, the results of which can take a long time to converge. However, the model is well suited to parallelisation and the effect of the disadvantage is reduced with the capabilities of modern digital computers.

When considering the challenges that will be faced in producing an asset management model that is capable of modelling the wide array of assets across the entire rail network and their associated inspection and maintenance, the PN model seems best suited to the task. The fact that PN modules can be constructed effectively in isolation from other modules means that a 'library' of modules can be constructed for the different assets in the network. According to the features of the network to be modelled, the appropriate modules could be taken from the library and used to construct a network PN model. This modularity extends also to other features such as the number and type of maintenance machines. Add to this the fact that neither the PN model nor the Monte Carlo simulation solution routine require the application of any restrictive assumptions, and an extended version of the PN model presented in reference 14 is extremely well suited to modelling a complex network of railway assets. Such a model could be used to investigate a combined approach to the management of many different railway assets. The model would be an invaluable tool to assist asset managers and maintenance decision makers in achieving the efficient and effective application of maintenance while ensuring minimal disruption of service. Whilst the Markov model is not as well suited to a network-wide model, its simplicity, both in terms of construction and analysis, make it a good tool for the rapid evaluation and investigation of individual network assets under specific sets of assumptions.

Acknowledgement

The authors gratefully acknowledge the support of Lloyd's Register Foundation * (LRF), the Royal Academy of Engineering and Network Rail.
* Lloyd's Register Foundation supports the advancement of engineering-related education, and funds research and development that enhances safety of life at sea, on land and in the air.

References

1. Shenton M.J., Ballast deformation and track deterioration, Track Technology, Tomas Telford Ltd, p253-264, 1985.
2. Chrismer, S. and Selig, E.T., Computer model for ballast maintenance planning, Proc 5th Int heavy haul railway conf, Beijing, p223-227, 1993.

3. Sato, Y., Optimum track structure considering deterioration in ballasted track, Proc of int heavy haul conf, Cape Town, South Africa, p576-590, 1997.
4. Hamid, A. and Gross, A., Track-quality indices and track degradation models for maintenance-of-way planning, Transportation research Board, 802, p2-8,1981.
5. Bing, A.J. and Gross A., Development of railway track degradation models, Transportation research Board, 939, p27-31, 1983.
6. Sadehghi, J. And Askarinejad, H., 'Development of Improved Railway Track Degradation Models', Structure and Infrastructure Engineering, Vol 6, No 6, p675-688,2010.
7. Shafahi, Y. and Hakhamameshi, R., Application of maintenance management model based on Markov chain and probabilistic dynamic programming for the Iranian railways, Trans A: Civil Engineering, Vol 16, No 1, p87-97, 2009.
8. Lyngby N., Hokstad, P. and Vatn, J., RAMS Management of Railway Tracks, published in the Handbook of Performability Engineering (Ed Krishna Misra), Springer Press, 2008, p1123-1145.
9. Podofillini, L., Zio, E., and Vatn, J., Risk-informed optimisation of railway tracks inspection and maintenance procedures, Journal of Reliability Engineering and System Safety 2006, 91, p20–35.
10. Kumar, S., Espling, U., and Kumar, U., A holistic procedure for rail maintenance in Sweden, Proc. of the IMechE, Part F: J. Rail and Rapid Transit 2008, 222 (4), p331-344.
11. Quiroga, L.M., and Schnieder, E., 'Monte Carlo Simulation of Railway Track Geometry Deterioration and Restoration', Proc. of the IMechE, Part O: J. Risk and Reliability, June 2012, 226(3), pp274-282.
12. Quiroga, L.M., and Schnieder, E., 'A Heuristic Approach to Railway Track Maintenance Scheduling', WIT Transactions on The Built Environment, Vol 114, p687-699, 2010.
13. Andrews, J.D., 'A modelling approach to railway track asset management,' Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 0954409712452235, first published on July 13, 2012 as doi:10.1177/0954409712452235
14. Prescott, D.R., and Andrews, J.D., 'A Railway Track Ballast Maintenance and Inspection Model for a Rail Network', Accepted for publication in Proc. of the IMechE, Part O: J. Risk and Reliability.
15. Prescott, D.R., and Andrews, J.D., 'Modelling maintenance in railway infrastructure management', In Proceedings of RAMS 2013 (Reliability, Availability and Maintainability Symposium), Orlando, USA, January 2013, [CD-ROM].
16. Billinton, R. and Allan, R.N., 'Reliability Evaluation of Engineering Systems: Concepts and Techniques,' 2nd Ed. New York: Plenum Press, 1992.
17. Andrews JD. 'Railway Track Ballast Condition Monitoring,' Proc. of ESReDA Seminar: Advances in Reliability-Based Maintenance Policies, La Rochelle, France, 5-6 Oct 2011.

A simple model of the software failure rate

Hendrik Schäbe
TÜV Rheinland InterTraffic GmbH, Germany

Abstract

In the paper a very simple model of a software failure rate is derived. This model is not intended to compute the failure rate or failure probability of software under realistic conditions. It is rather used to show the influences of measures for software quality assurance and of measures required by software safety standards on this failure rate, i.e. a tendency. In particular, it is shown that testing is important to reduce software failure rate and the effect is the larger, the closer testing is to exhaustive testing. Verification is also important and it needs to be done by an independent person or party to give a good effect. Another result is that software failure rate grows with the size of the software, the frequency the software is used and, the fraction of the software that is used per demand cycle. As a result, it is recommended to implement the measures with more consequence and rigidity, the larger the part of the software that is used, the larger the software itself and the more frequent it is used.

1. Introduction

This paper is dedicated to the derivation of a very simple model of a software failure rate. This model is not intended to compute the failure rate or failure probability of software under realistic conditions. It will be used to show the influences of measures for software quality assurance and measures required by software safety standards on this failure rate, i.e. a tendency.

The frequently used reliability growth models (e.g. Musa et al. (1987)) or other models as e.g. Bayesian Belief networks (Kang et al. (2012), Holmberg et al. (2012)) that are intended to statistically estimate software failure probability will not be used. These traditional models give no direct connection between human errors in the software process and the resulting software failure rate. Therefore, a very simple model that links human error to software failures is presented here. The model has not been used before. The model is not intended to compute a software failure rate but rather to show tendencies. In section two the model is presented. The third section shows some influences of verification, testing etc. on the software failure rate.

2. A model of the software failure rate

2.1 The model

Failures of the software that is part of a software system occur randomly. This, however, is a consequence of the use and the environment of the software, which generate random input to the software. If the software is not able to reconfigure, it is static and a failure of a system with software is always caused by the fact that a software error, i.e. a deviation from the intended

function, is activated. The software error has been present from the very beginning. So, it depends on the circumstances, when the software error is activated and leads to a system failure, Kang (2012).

Software is usually produced by a process like the following.

1. Software requirements specification

   A software requirements specification is elaborated based on the user requirements. These can be part of a system requirement specification.

2. Software architecture

   Based on the software requirements the architecture of the software is specified. This is the structure itself but also the requirements for the parts, e.g. the modules. Note that, this step might be iterated if e.g. the software is divided into parts, which are divided into component and those further into modules etc. We will only consider a one step software architecture.

3. Coding

   In this phase, the code itself is produced based on the description of the smallest software architecture unit defined in the phase above.

Note that, the phases of software integration are taken into account later, when we discuss the measures for software quality improvement.

Now, in each of these three phases a different amount of work can be carried out. This can be described by

1. The number of requirements n_1
2. The number of architectural requirements (module requirements) n_2
3. The number of statements n_3.

One observes that

$n_1<n_2<n_3$.

It can be assumed that an error can occur during the three phases with the coinciding probabilities p_1, p_2, and p_3. Since the complexity of the tasks is decreasing starting from phase 1, we have

$p_1>p_2>p_3$,

see e.g. Dhillon (1986) on this issue. If N denotes the number of software errors, then it has mean

$$E(N) = n_1p_1+n_2p_2+n_3p_3 \quad (1)$$

and variance

$$Var(N) = n_1p_1(1-p_1)+n_2p_2(1-p_2)+n_3p_3(1-p_3) \quad (2)$$

and the number N will be distributed asymptotically normally for large n_1, n_2, n_3, i.e. for large software. This result is derived from the fact that N is a sum of a large number of random variables. It has been assumed that the

environment of the software behaves sufficiently in a random manner to trigger the software errors in a random manner.

In order to come to a software failure rate one must take into account:

a) Software is often used in demand cycles, i.e. a demand occurs which triggers the software to work. At the end of the cycle, the task is finished and the software waits for the next demand. Let ν denote the demand rate.

b) Not the entire software might be used, but only a part of it. Let A describe the average fraction of the software that is used per demand.

Then, the software failure rate is

$$\lambda = NA\nu. \qquad (3)$$

Here, it has been assumed that the probability of an error per line of code is almost constant throughout the software and that random parts of the software are activated, depending on the (random) input values generated by the environment.

In this model, the lifetime of the software, i.e. the time to first failure will have an exponential distribution. Note that, it will be conditionally on N, because the number of errors in the software is fixed for a certain software release. If after failure the software is restarted, i.e. “repaired” and restored into a status as good or bad as old, then the failures of the software system will occur according to a homogenous Poisson process, i.e. with exponentially distributed time intervals with rate λ.

When studying (3) one can make the following observations:

a) failure rate of the software is proportional to the size,

b) failure rate of the software is proportional to the fraction of the software used during one cycle,

c) failure rate of the software λ is proportional to the number of software errors N, which is trivial.

2.2 Simple example

A simple numeric example can be computed as follows.

Assume:

p_1=0.1, p_2=0.01, p_3=0.001

to reflect different complexity of tasks and the coinciding human error probabilities (see B.S. Dhillon to support these rough estimates),

n_1= 1000, n_2= 10000, n_3=100000.

One obtains

E(N) = 100+100+100=300.

If 10% of the software is used during one cycle (A=0.1) which is assumed to take an hour, this yields

λ = 30/h.

The derived failure rate is very large. This is caused by the fact that

a) the software is quite complex (100000 statements)

b) no measures in the software process as testing, verification, structured design etc. have been taken into account.

3. Modelling verification, testing and other methods

In this section simple models for testing, verification and other methods will be presented that can be applied in the software process. With the help of these models the influence of these measures on software failure rate is studied.

3.1 Testing

Assume the software is tested and that the test is systematic, i.e. it covers the requirements of the software requirements specification, the architecture and the coding. If the test is exhaustive, all software errors would be revealed and the software would be error-free – with respect to the specification against which it is tested. This, however, is almost impossible.

In practice, there would be coverages C_1, C_2 and C_3 that show, which part of the requirements and / or statements are covered by tests in the three phases. Then, the number of errors N has a modified mean

$$E(N) = n_1p_1(1-C_1)+n_2p_2(1-C_2)+n_3p_3(1-C_3) \quad (4)$$

and variance

$$Var(N) = n_1p_1(1-C_1)\,(1-p_1(1-C_1))+n_2\,p_2(1-C_2)\,(1-p_2(1-C_2))+n_3p_3\,(1-C_3)\,(1-p_3(1-C_3)) \quad (5)$$

Again, the number of software errors is asymptotically normally distributed.

Assume now the following test coverages

C_1=1, C_2=0.999 and C_3=0.99

and one arrives at

E(N) = 0.1+1 =1.1

and hence at a failure rate of

λ=0.11/h,

which is significantly smaller.

3.2 Verification

Now, verification is introduced into the model, i.e. the results of a phase are crosschecked against the output of the previous phase. That means, the code can be verified against the architectural description, the architectural

description can be checked against the software requirements specification and the software requirements specification against the system requirements specification. Let V_1, V_2 and V_3 denote the coverage of the verification steps. The coverage shall denote simply the probability that an existing error is detected during verification and removed afterwards. In fact, it is the combination of the part of the software that is verified and the probability that a failure is detected during verification. Now, formulae (4) and (5) need to be modified. This gives

$$E(N) = n_1p_1(1-C_1)\,(1-V_1)+n_2p_2(1-C_2)\,(1-V_2)+n_3p_3(1-C_3)\,(1-V_3) \qquad (6)$$

and

$$\begin{aligned} Var(N) = {} & n_1p_1(1-C_1)(1-V_1)(1-\,p_1(1-C_1)(1-V_1) \\ & + n_2p_2(1-C_2)(1-V_2)(1-\,p_2(1-C_2)(1-V_2) \\ & + n_3p_3(1-C_3)(1-V_3)(1-\,p_3(1-C_3)(1-V_3) \end{aligned} \qquad (7)$$

Here it is assumed that verification is independent from the process itself, i.e. it is done by an independent person or team.

Assume that $V_1=V_2=V_3=0.999$. Then, in this example

$$E(N) = 0.1*0.001 + 1*0.001 = 1.1\ 10^{-3}$$

which gives

$$\lambda = 1.1\ 10^{-4}/h.$$

This value is already and closer to realistic values for software.

If now verification is not independent from software development one cannot assume that the failure probability is

$$p(1-C)(1-V).$$

It is a value that is larger than this, but smaller than p(1-C). For simplicity assume that dependent verification leads to a failure probability of

$$P(1-C)\sqrt{1-V}\ .$$

This assumption has been made for computational purposes and is no result on error detection in dependent verification.

In the example this gives the result

$$E(N) = 0{,}00316 +0{,}0316 = 0{,}0348$$

yielding

$$\lambda=0.00348/h,$$

which is smaller than without verification but larger than with independent verification.

3.3 Failures in the software specification

When considering the specification phase (first phase) of the software process one has to admit that testing and verification is always testing or verifying something against another thing. The software requirements specification is verified against the intention of the user and the software is tested to comply with the user requirements. If, however the user requirements which form the starting point are (at last) partially wrong and the rest of the software process is correct, i.e. it coincides with the user requirements –then the result is (partially) wrong. All other errors can be (theoretically) detected by testing and verification, however the starting point is hard to test and verify. Moreover, even the probabilities for verification and testing at the level of the software requirements specification are larger and test coverages are smaller than in the other phases. For the later phases, one can achieve very good test coverages and very good error detection via verification, e.g. using validated and proven tool chains. Hence, for a software that is a result of a good software process, the errors would mainly be in the specifications: mainly the user requirements and the software requirements specification, see Malm et al. (2012).

4. Conclusions

A very simple model for a software failure rate has been constructed, which is not intended to compute a rate but to show the influence of measures in the process. It has been shown that:

- Testing is important to reduce software failure rate and the effect is the larger, the closer we come to exhaustive testing.
- Verification is also important and it needs to be done by an independent person or party to give a good effect.

Moreover it turned out that software failure rate grows with

- the size of the software,
- the frequency the software is used and,
- the fraction of the software that is used per demand cycle.

The latter aspects are not taken into account when setting up requirements for safety related software as e.g. in IEC 61508. Therefore, it is desirable to implement the measures with more consequence and rigidity, the larger the part of the software that is used, the larger the software itself and the more frequent it is used.

References

1. B. S. Dhillon, Human reliability, Pergamon Press, 1986
2. IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems Part 3: Software requirements, 2010

3. J.-E. Holmberg, P. Bishop, S. Guerra, N. Thuy, Safety case framework to provide justifiable reliability numbers for software systems , 11th International Probabilistic Safety Assessment and Management Conference, The Annual European Safety and Reliability Conference, Helsinki, Finland 25–29 June 2012, Paper _10-Th2-2_-_Holmberg

4. H.G. Kang, H.-S. Eom, H.S. Son, S.C. Jang, , An Integrated Quantification Method for Safety-Critical Software Failure Probability , 11th International Probabilistic Safety Assessment and Management Conference, The Annual European Safety and Reliability Conference, Helsinki, Finland 25–29 June 2012, paper _10-Th2-1_-_Kang

5. Safety-Critical Software Failure Probability, 11th International Probabilistic Safety Assessment and Management Conference, The Annual European Safety and Reliability Conference, Helsinki, Finland 25–29 June 2012, , Paper _10-Th2-1_-_Kang

6. M. Kristiansen, R. Winther, B. Natvig,, A component-based approach for assessing reliability of compound software, 11th International Probabilistic Safety Assessment and Management Conference, The Annual European Safety and Reliability Conference, Helsinki, Finland 25–29 June 2012, , Paper _10-Th2-3_-_Kristiansen

7. T. Malm, M. Hietikko, and J. Rauhamäki, Comparing safety requirement sources of machinery software, 11th International Probabilistic Safety Assessment and Management Conference, The Annual European Safety and Reliability Conference, Helsinki, Finland 25–29 June 2012, , Paper _10-Th3-1_-_Malm

8. J.D. Musa, A. Iannino, K. Okumoto, Software reliability, McGraw Hill 1987

PUBLISHED BY:
Loughborough University, Loughborough, Leicestershire, LE11 3TU.

FOR FURTHER INFORMATION CONTACT:
Kate Sanderson, Nottingham Transportation Engineering Centre, University of Nottingham
University Park, Nottingham, NG7 2RD

Tel: +44 (0)115 9513953 Email: Kathryn.Sanderson@nottingham.ac.uk

ISBN 978 1 907382 61 1