On the challenge of designing a robust military force

A multi-resolution modelling approach to improve the performance of a naval force support system

S.R.M. Salome

© *iStock*

Master thesis submitted to Delft University of Technology

In fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Engineering & Policy Analysis**

Faculty of Technology, Policy and Management

by

S.R.M. Salome

Student number: 4575768

Defended in public on November 16, 2021

Graduation committee

Chairperson:	Prof. dr. ir. A. Verbraeck, Policy Analysis
First Supervisor:	Dr. ir. W. L. Auping, Policy Analysis
Second Supervisor:	Dr. ir. Y. Huang, Systems Engineering
First External Supervisor:	Drs. A. Ros, Copernicos
Second External Supervisor:	Ir. M. M. R. Kuijer, Copernicos

(this page is intentionally left blank)

-DISCLAIMERS-

[DISCLAIMER 1] This research is conducted in collaboration with Copernicos. Copernicos is a tech consultancy with a strong focus on both practical and scientific innovation in the asset management field. The case used in this research involves a simulated model developed by Copernicos for the Dutch Royal Navy. This simulation model reflects certain elements of the design of the current force support system of the Dutch Royal Navy. However, as models are a simplification of the real world, the findings in this research can under no circumstances be generalized to the real-world performance of Dutch military vessels with respect to their availability and costs. In this light, this research should be not be read as an evaluation of the Dutch military capability but rather as a scientific study that evaluates the use of a multi-resolution modelling method.

[DISCLAIMER 2] As the WSMD model contains many classified content about the performance of Dutch military vessels, some parts of this research remain confidential. More concrete, this thesis report does not include the operational data related to military vessels, the aggregation functions, the specification of the aggregated model, the datasets associated to exploratory analysis, and the practical research recommendations. Unfortunately, this implies that this research will not reproducible. Still, it is believed that the confidentiality of this research does not affect the scientific conclusions.

(this page is intentionally left blank)

Preface

This thesis concludes my time as a student at the faculty of Technology, Policy and Management at the TU Delft. Education-wise, the past years have been inspiring. What truly fascinates me is the intersection of the modelling & simulation field and the decision-making field. History has demonstrated many cases where decision-makers misuse simulation models by mistakenly treating them as a holy grail for policy design. This emphasizes the large responsibility of system engineers for their clear communication of the purpose of a model, which inherently comes with limitations on its use. After all, as the great statistician George Box once wrote: "all models are wrong, but some are useful".

Apart from the great education, this program also brought me some valuable friendships. In particular, I would like to mention the friendship with Romek van Deursen, who I met at the first day of the bachelor's study program. He passed away one year ago, and I would like to dedicate this thesis to him.

I most certainly could not have done this study effort alone. A special thanks to my external supervisors Arjen Ros and Michel Kuijer from Copernicos, who made this research possible by providing me a simulation model currently in use by the Dutch Royal Navy. They have also put much effort in guiding me throughout the simulation model in question, for which I am thankful. Another special thanks to my TU Delft first supervisor Willem Auping for his substantive support throughout my thesis. I would also like to thank the other TU Delft committee members Alexander Verbraeck and Yilin Huang for giving me helpful feedback during my kick-off meeting, mid-term meeting, and green-light meeting.

I would also like to thank people in my social circle. First of all, I count myself very lucky with my fiancée. I am grateful for her mental support and her unconditional faith in my ability to conduct this research. Next, I would like to thank my parents for their mental and financial support. Last but not least, I would like to thank my friends and roommates for all the fun and relaxing moments throughout my study.

I realize myself that I have been very privileged to follow this education and experience the student life. Now, with all the things that I have learned, I hope that I can create value in the society we live in.

I hope you enjoy reading my thesis.

Stefan Salome
Delft, November 2021

(this page is intentionally left blank)

Executive summary

In order to ensure a nation's safety & security, it is of the highest importance that military forces convey a capability that can counter multiple threats. However, given limited budgets, defence organizations need to make choices regarding the composition of a force and the extent to which a force can be kept deployable. In order to get insight in the consequences of such choices, quantitative models are used to get insight in the force capability under different future conditions. Often, detailed models are developed to mimic complex phenomena. However, due to high analytical and computational costs, such models tend to be 'bad' in the evaluation of uncertainty in a system. This can be problematic, as the outcomes of detailed models usually depend on numerous dubious assumptions that are too buried to be understood by hand.

In this light, there has always been a clear need to have models with a *low* resolution next to models with a *high* resolution. 'Model resolution' is defined as *"the degree of detail used to represent aspects of the real world or a specified standard or referent by a model or simulation"* (Defense Modeling & Simulation Enterprise, n.d.). As low-resolution models incorporate less detail, they are better able to evaluate the effect of uncertainty in a system. In this light, low-resolution models also facilitate long-term reasoning which is important for strategic policy design.

Now, although the combined use of simple models and detailed models is potentially promising, few practical studies exist that systematically work with models on different resolutions. In order to narrow this research gap, this study examines the use of models at different resolutions to improve the robustness of a naval force support system. A force support system is an umbrella term for all supporting systems that together determine the force availability; e.g. maintenance systems, reliability systems, etc. Robustness refers to some measure of insensitivity of a strategy across a wide range of scenarios. In this study, the performance of a naval force support system is measured with the sailing availability of a vessel and the working hours of its crew.

In order to work with models at different resolutions, an existing detailed model that reflects certain elements of the force support system of the Dutch Royal Navy is used to develop a new 'simpler' model. This practice of moving from a high resolution to a low resolution is called 'aggregation'. The detailed model (hereinafter referred to as the Weapon Systems Management Dynamics (WSMD) model) mimics the availability of a military vessel over its lifetime, based on a detailed pre-specified mission schedule. In theory, the simple model (hereinafter referred to as the aggregated model) can extend the usefulness of the WSMD model by exploring the rough availability of a military vessel over a wide range of possible scenarios.

Although the flexibility of the aggregated model would ideally complement the complexity of the WSMD model, history has demonstrated many cases where decision-makers dismiss low-resolution models due their inability to capture critical 'complex' information. After all, moving from complexity to simplicity eminently comes with a loss of information. In this light, it is the question how a 'valid' low-resolution model can be developed and used to improve the robustness of a naval force support system. More general, this study aims to answer the following main question:

"How can multi-resolution modelling be useful in improving the robustness of a naval force support system"?

The main conclusion of this study is that multi-resolution modelling can be considered useful as the detailed WSMD model and the simplified aggregated model were found to be mutually dependent in their design, validity, and use to improve the robustness of a naval force support system Table S.1 summarizes the findings that substantiate this conclusion.

Table S.1. Main findings that substantiate the mutual dependency of the WSMD model and the aggregated model

	Design	Validity	Use
WSMD model → AGG model	Together with a set of trivial and non-trivial aggregation functions, the WSMD model formed the basis for the design of an aggregated model.	Together with a set of cross-validation tests, the WSMD model formed the basis for the establishment of the fitness for purpose of the aggregated model.	As real-world policies usually require a specification on a detailed level, the complexity of the WSMD model remains important for the evaluation of the effectiveness of promising actions discovered by the aggregated model.
AGG model → WSMD model	Via the validation process, the aggregated model had design implications regarding some potentially problematic assumptions of the WSMD model.	By generating many scenarios, the aggregated model revealed some potentially problematic assumptions in the WSMD model. This may decrease the validity of the WSMD model.	The exploratory use of the aggregated model allows decision-makers to get insight in problematic scenarios and strategic trade-offs. In particular, the aggregated model revealed robustness patterns that are hard to see through with a few detailed scenarios of WSMD model.

The following three paragraphs provide a more elaborate description of each leg.

Design

With the use of aggregation functions, the numerous inputs of the detailed WSMD model were mapped to only a few inputs of the aggregated model. Many types of aggregation functions were implemented, ranging from understandable functions such as the mean and the sum to entirely different modelling structures. The choice for the type of aggregations depended on the degree of non-linearity and object heterogeneity in the WSMD model. In fact, the main conclusion of the design phase is that glossing over qualitatively different phenomena and objects likely results in a poor aggregation. In case that simple aggregation functions were inadequate, this study 'approximated' most low-level information by (a) adopting a new modelling structure, (b) creating 'abstract' object groups, and/or (c) introducing a calibration vector. Such approximations involved trade-offs between the computational costs, the interpretability, and the consistency error of the aggregated model.

Validity

Model validation was considered a two-way process as both models could have implications for each other's validity. The cross-validation tests measured the fitness for purpose of the aggregated model by comparing its outcomes to the outcomes of the WSMD model. In fact, treating the aggregated model in isolation would obscure its merit as it would be unclear in what cases it sufficiently reflects the actual dynamics of a force support system. Given the cross-validation tests, no problematic information loss in the aggregated model was encountered. Vice versa, with an ensemble of scenarios, the aggregated model was able to validate 'hidden' assumptions in the WSMD model. In general, some static assumptions in the WSMD model may need replacement by dynamic balancing feedbacks that provide a better representation of real-world dynamics. Without such feedbacks, defence decision-makers may push the system in the wrong direction when trying to eliminate a performance gap.

Use

The exploratory use of the aggregated model gave insight in the relation between the robustness of the naval force support system and the number of large maintenance (or overhaul) operations in the lifetime of a vessel. In fact, a linear increase in the time between two consequent overhauls exponentially decreases the robustness of the current maintenance plans as depicted by the models. The main driver of this relation is the exponential degradation behaviour of a vessel due to lower-level interactions between objects. Especially in case of a scenario characterized by a high degradation, a long overhaul interval, and/or a low crew productivity, the robustness of the force support system significantly decreases with regard to its sailing availability and working effort. Based on these scenarios, the aggregated model also found a number of leverage points to increase the robustness of the force support system. Still, the WSMD model remains crucial for the real-world implementation of promising actions discovered by the aggregated model. As high-resolution models can better represent real-world complexity, they can better assess the actual effectiveness of policies.

Limitations and scientific research recommendations

This study comes with a number of limitations and scientific research recommendations. The most important limitation of this study concerns the fact that the aggregated model could not be fully cross-validated with the WSMD model. This hindered the full establishment of the fitness of the purpose of the aggregated model. Other limitations concern the narrow scope of aggregated model and the way in which high-level policies are specified. Scientific research recommendations concern an improved cross-validation process that provides a more thorough definition of consistency and the development and evaluation of disaggregation functions. Furthermore, it is strongly recommended to research the composability of MRM in the future. Composable models at different resolutions allow practitioners to efficiently address questions at different levels of detail in a certain area of interest.

Management samenvatting

Om een land veilig te houden, moeten militaire organisaties laten blijken ze dat een verscheidenheid aan dreigingen kunnen weerstaan. Echter zijn de budgetten gelimiteerd, daarom moeten organisaties keuzes maken in de samenstelling van een krijgsmacht en haar mate van inzetbaarheid. Veel militaire organisaties gebruiken kwantitatieve modellen om inzicht te krijgen in de gevolgen van verschillende keuzes onder verschillende omstandigheden. Vaak worden gedetailleerde modellen gebruikt om complexe verschijnselen na te bootsen. Echter, om verschillende analytische en rekenkundige redenen zijn zulke modellen gelimiteerd in het adresseren van onzekerheid in een systeem. Dit kan problematisch zijn, omdat de uitkomsten van gedetailleerde modellen vaak afhankelijk zijn van vele 'twijfelachtige' aannames die te verborgen zijn om handmatig te doorgronden.

Om deze reden is er altijd vraag geweest naar modellen met een *lage* resolutie naast modellen met een *hoge* resolutie. 'Model resolutie' is gedefinieerd als *"de mate van detail om aspecten van de wereld te representeren met een model of simulatie"* (Defense Modeling & Simulation Enterprise, n.d.). Omdat lage-resolutie modellen minder detail bevatten, zijn ze beter in het evalueren van het effect van onzekerheid in een systeem. Hierdoor kunnen lage-resolutie modellen ook het lange-termijn denken faciliteren, wat belangrijk is voor het ontwikkelen van strategisch beleid.

Hoewel het gebruik van zowel simpele als gedetailleerde modellen veelbelovend is, bestaan er weinig praktische studies die systematisch werken met modellen op verschillende resoluties. Om dit kennistekort te adresseren, onderzoekt deze studie het gebruik van modellen op verschillende resoluties om de robuustheid van een maritiem ondersteuningssysteem te verbeteren. Een ondersteuningssysteem van een krijgsmacht is een overkoepelende term voor alle systemen die gezamenlijk de inzetbaarheid van een krijgsmacht bepalen; bijvoorbeeld onderhoudssystemen, betrouwbaarheidssystemen etc.). Robuustheid refereert naar de mate waarin een strategie ongevoelig is voor een grote verscheidenheid aan scenario's. In deze studie is de prestatie van een maritiem ondersteuningssysteem gemeten met de beschikbare vaartijd van het vaartuig en de werkuren van de bemanning.

Om met modellen op verschillende resoluties te werken, wordt een bestaand gedetailleerd model gebruikt dat verschillende elementen van het ondersteuningssysteem van de Koninklijke Nederlandse Marine representeert. Dit gedetailleerde model (in het vervolg het WSMD model genoemd) wordt gebruikt om een nieuw, simpeler model te ontwikkelen. Het WSMD model bootst de beschikbaarheid van een militair vaartuig na, gebaseerd op een vooraf gespecificeerd missieprofiel. In theorie kan het simpele model (in het vervolg het geaggregeerde model genoemd) de bijdrage van het WSMD model vergroten door de beschikbaarheid van een militair vaartuig te verkennen aan hand van vele mogelijke scenario's.

Hoewel de flexibiliteit van het geaggregeerde model de complexiteit van het WSMD model kan completeren, is het al vaak voorgekomen dat besluitvormers het gebruik van lage-resolutie modellen afwijzen vanwege hun gebrek om cruciale informatie voor besluitvorming te bevatten. Immers, het omzetten van complexiteit in simpliciteit komt inherent met een verlies aan informatie. Om deze reden is het de vraag hoe een valide geaggregeerd model ontwikkelt en gebruikt kan worden om de robuustheid van een maritiem ondersteuningssysteem te verbeteren. Meer algemeen benaderd probeert deze studie de volgende hoofdvraag te beantwoorden:

"Hoe kan multi-resolutie modellering nuttig zijn in het verbeteren van de robuustheid van een maritiem ondersteuningssysteem?

De belangrijkste conclusie is dat multi-resolutie modellering nuttig is omdat het gedetailleerde WSMD model en het simpele geaggregeerde model wederzijds afhankelijk zijn in hun ontwikkeling, validiteit, en gebruik om de robuustheid van het maritieme ondersteuningssysteem te verbeteren. Tabel S.1 vat de bevindingen samen.

Tabel S.1. Belangrijkste bevindingen die de wederzijdse afhankelijkheid van het WSMD model en het geaggregeerde model illustreren.

	Ontwikkeling	**Validiteit**	**Gebruik**
WSMD model → AGG model	Samen met een aantal triviale en niet triviale aggregatie functies vormt het WSMD model de basis voor de ontwikkeling van het geaggregeerde model.	Samen met een aantal cross-validatie testen, vormt het WSMD model de basis voor het bepalen de validiteit van het geaggregeerde model.	Omdat de implementatie van beleid vaak op gedetailleerd niveau is, blijft de complexiteit van het WSMD model belangrijk voor de evaluatie van de effectiviteit van aanbevolen acties door het geaggregeerde model.
AGG model → WSMD model	Via het validatie proces heeft het geaggregeerde model implicaties voor de modellering van potentiële problematische aannames in het WSMD model	Door veel scenario's te genereren, onthulde het geaggregeerde model een aantal potentieel problematische aannames in het WSMD model. Deze aannames kunnen de validiteit van het WSMD model verlagen.	Het exploratieve gebruik van het geaggregeerde model geeft besluitvormers inzicht in problematische scenario's en strategische afwegingen. Het geaggregeerde model onthulde bijv. robuustheidspatronen die moeilijk inzichtelijk te krijgen zijn met het WSMD model.

De volgende drie alinea's wijden verder uit over bevindingen in tabel S.1.

Ontwikkeling

De vele invoer parameters in het WSMD model zijn met behulp van aggregatie functies getransformeerd naar enkele invoer parameters in het geaggregeerde model. Veel verschillende aggregatie functies zijn geïmplementeerd, variërend van simpele functies zoals het gemiddelde en de som tot volledig nieuwe modelleerstructuren. De keuze tussen verschillende typen aggregatie functies was afhankelijk van de mate van niet-lineariteit en object heterogeniteit in het WSMD model. In feite was de belangrijkste conclusie van de ontwikkelingsfase dat het aggregeren van kwalitatief verschillende fenomenen en objecten tot verkeerde benaderingen op een hogere resolutie leidt. In het geval dat simpele aggregatie functies niet geschikt waren, werd gedetailleerde informatie in het WSMD model benaderd met (a) nieuwe modelleer structuren, (b) abstracte object groepen, en/of (c) kalibratie vectoren. Zulke benaderingen kwamen met onvermijdelijke afwegingen tussen de rekenkundige complexiteit, de interpreteerbaarheid, en de consistentie fout van het geaggregeerde model.

Validiteit

Model validatie was een tweezijdig proces omdat beide modellen de validiteit van elkaar bepalen. De validiteit van het geaggregeerde model werd bepaald door cross-validatie testen die de uitkomsten van het geaggregeerde model vergeleken met de uitkomsten van het WSMD model. Zonder het WSMD model en de cross-validatie testen zou het onduidelijk zijn in hoeverre het geaggregeerde model bruikbaar is voor een analyse. Met de cross-validatie testen werd geen problematisch informatie verlies in het geaggregeerde model gevonden. Omgekeerd kon het geaggregeerde model verborgen aannames van het WSMD model valideren door ze op een verscheidenheid aan scenario's te testen. Hierdoor moeten statische aannames in het WSMD model mogelijk vervangen worden door meer dynamische balancerende feedback structuren die een betere representatie zijn van de werkelijke dynamiek in een maritiem ondersteuningssysteem. Zonder zulke dynamische structuren kunnen besluitvormers verkeerde beslissingen nemen om de prestatie van een zeemacht te verbeteren.

Gebruik

Het exploratieve gebruik van het geaggregeerde model gaf inzicht in de relatie tussen de robuustheid van het maritieme ondersteuningssysteem en het aantal keren groot onderhoud in de levensduur van een vaartuig. Een belangrijke bevinding was dat een lineaire verlenging van de tijd tussen twee 'groot onderhoud' operaties leidde tot een exponentieel verval van de robuustheid van de huidige onderhoudsplannen zoals aangeduid in de modellen. Dit wordt verklaard door de exponentiele degradatie van een asset, te wijten aan interacties tussen objecten op een lager niveau. De robuustheid van een onderhoudssysteem werd met name verlaagd door scenario's met (a) een hoge natuurlijke degradatie van objecten, (b) een lange tijd tussen twee 'groot onderhoud' operaties en/of (c) een lage productiviteit in onderhoud van de bemanning. Deze scenario's belichtten tegelijkertijd ook potentieel invloedrijke acties in het geaggregeerde model waarmee de robuustheid van het ondersteuningssysteem verhoogd kon worden. Echter blijft het WSMD model belangrijk voor het evalueren van het beleid dat als veelbelovend wordt beschouwd door het geaggregeerde model. Hoge-resolutie modellen kunnen immers beter de werkelijke complexiteit van een ondersteuningssysteem nabootsen, en kunnen daarom beter de daadwerkelijke effectiviteit van beleid evalueren.

Limitaties en wetenschappelijke aanbevelingen voor onderzoek

De belangrijkste limitatie van dit onderzoek betreft het feit dat het geaggregeerde model niet volledig ge-cross-valideerd kon worden met het WSMD model. Hierdoor was het lastig om de volledige validiteit van het geaggregeerde model vast te stellen. Andere limitaties betreffen de kleine scope van het geaggregeerde model en de manier waarop beleid is gespecificeerd. Wetenschappelijke aanbevelingen pleiten onder andere voor een verbeterd cross-validatie proces dat voorziet van een meer grondige definitie van model consistentie. Ook het gebruik van disaggregatie functies in verder onderzoek wordt aanbevolen. Tot slot wordt aanbevolen om meer onderzoek te doen naar de combineerbaarheid van modellen op verschillende resoluties. Combineerbare modellen staan besluitvormers toe om vraagstukken op een efficiënte manier op verschillende detail niveaus te beantwoorden.

Table of contents

List of figures

List of tables

List of equations

(this page is intentionally left blank)

1. Introduction

Defence planning is critical for the military capability of defence organizations (Gray, 2014). In defence jargon, military capability refers to the ability to perform military actions to achieve desired objectives or effects in a specific operating environment (Young, 2015). Nowadays, 'capability-based planning' (CBP), which was a revolution for defence planners two decades ago, has become the norm for NATO members (De Spiegeleire, 2011). The essence of CBP is to start by assessing a wide range of possible future military needs and then to develop a force that is able to satisfy that (Davis & Finch, 1993). Another essential aspect of CBP is that the military capability is made up of the combined effect of multiple elements, which usually includes training, equipment and personnel but also organisation, leadership and doctrine (Oxenham, 2010). CBP was presented as a huge improvement, as the previous planning strategies were more rigid; less likely trends were discarded and links between different capability elements were often left unnoticed (Oxenham, 2010; Davis, 2018). Surprisingly however, there are still serious concerns about the poor inclusion of uncertainty in current defence planning strategies of NATO members (De Spiegeleire, 2011; Burk & Parnell, 2011; Navarro-Galera et al., 2011; Tate & Thompson, 2017; Davis, 2018).

The lack of uncertainty considerations in defence planning is problematic. To justify, it is of the highest importance that a force conveys a robust capability that can counter multiple threats (NATO, 2020). As robustness refers to some measure of insensitivity of a strategy across the scenario space (Maier et al., 2016), excluding uncertainty considerations in defence planning processes inhibits the ability to design a force with a robust capability. This is not only problematic for the NATO member in question, but threatens the safety & security of the entire alliance (NATO, 2020). Now, which uncertainties matter for defence organizations when aiming for a robust force?

In defence planning, defence organizations need to deal with uncertainty in (a) the planning of force requirements and (b) the planning of force support (Wojtaszek & Wesolkowski, 2012). Although both planning types are interlinked and overlap in the uncertainty they deal with (McLucas, 2011), they are conceptually different. First, force requirement planning refers to what tasks will need to be carried out in the future and by which assets. This is a major uncertainty as the 21st century requires to deal with very different, possibly simultaneously occurring kinds of missions (Wesolkowski & Eisler, 2015; Davis, 2018). To illustrate, mission types typically include grey-war tactics due to geopolitical rivalry, peace operations, counterterrorism, and emergency assistance in the face of natural disasters (NATO, 2018). Second, force support planning refers to all supporting systems that determine the availability of a force. Force support systems are complex as they involve many interdependent subsystems such as maintenance systems, crew training systems and asset reliability systems (Adamides et al., 2004; Oxenham, 2010; Turan et al., 2020). These systems contain many uncertain factors; a minor change of which can have enormous consequences for the performance of a force (Bender et al., 2009; McLucas, 2011). Now that the problem of uncertainty is clear, the next step is a literature review in order to understand why uncertainty considerations are lacking.

As also noticed by Davis (2018), the majority of research regarding CBP relies on mathematical programming by (robustly) optimizing the force structure based on a variety of possible future missions and multiple objectives (Xiong et al., 2014; Shafi et al., 2017; Checco et al., 2017; Caron et al., 2019; Harrison et al., 2020). Some studies extend this with high-fidelity simulation in order to map complexities on a mission level (Eisler & Allen, 2012; Marlow & Novak, 2013; Wesolkowski & Eisler, 2015). However, both large optimization models and high-fidelity simulation tend to be bad in broadly assessing the impact of uncertainty as this usually results in an incomputable combinatorial complexity (Bankes, 1993). To illustrate, the duration of a run of the Canadian force design model 'Tyche' could already take up to several weeks (Eisler & Allen, 2012). However, as a consequence of not (or poorly) assessing the impact of uncertainty, defence decision-makers tend to give less weight to such models in their thinking as the outcomes depend on too many assumptions which are too numerous and buried to be easily understood (Davis, 2016).

Therefore, Davis (2014, 2016) advocates the development of simpler, 'higher-level' models in order to perform exploratory analysis and thereby increase model transparency and understanding. With Exploratory Modelling and Analysis (EMA), insights can be derived from the system's behaviour by displaying the pattern of policy performance over the entire uncertainty space of possible system models (Walker et al., 2013; Bankes et al., 2013). EMA is becoming increasingly popular for 'deeply' uncertain problems (Moallemi et al., 2020); i.e. problems where the future states of the system are unknown and/or decision-makers cannot agree upon them (Lempert et al., 2003; Maier et al., 2016). Regarding CBP, the use of EMA would allow defence decision-makers to design a robust force (Davis, 2003). Now, as CBP is increasingly referred to as deeply uncertain (Davis, 2018; Harrison et al., 2020), the number of exploratory studies that guide CBP is rising.

Exploratory studies can be found at any resolution. (Based on the working definition of the Defense Modeling & Simulation Enterprise (n.d.), this study defines resolution as *"the degree of detail used to represent aspects of the real world or a specified standard or referent by a model or simulation"*) Regarding objects, Bender et al. (2009), McLucas (2011), Malmi et al. (2011) and Zhang et al. (2014) model uncertainties regarding the availability or capability of equipment and single assets. On a fleet level, Abbass et al. (2008), Ahram et al. (2017), Moallemi et al. (2018), Elsawah et al. (2018), Ma (2019) and Turan et al. (2020) try to find effective acquisition policies in order to anticipate scenarios that can lead to an undesired availability or capability. One could move up the hierarchy of objects even more by focussing on the level of a national force and its role in a particular coalition (Gallagher et al., 2014). Exploratory studies not only differ in the resolution of objects; studies also adopt different resolutions in time, space (e.g. the detail in the mission area), and process (e.g. the detail in the maintenance process). However, although the use of EMA at a single resolution is promising, there is an increasing need to understand the relationships and influences of military systems that are represented at different resolutions (Loper & Register, 2015; Davis, 2016).

It is noticed that most studies either focus on detailed analysis in order to get insight in complex low-level phenomena, or on high-level analysis in order to facilitate exploratory analysis and (thereby) strategic reasoning. However, both high and low resolutions matter when one aims to understand the overall performance of a force. To clarify, it is sometimes needed to go into greater detail in case things are unclear, and to move upward in order to see the bigger picture (Davis & Hillestad, 1993; Davis & Bigelow, 1998; Bigelow & Davis, 2003). For instance, when one aims to understand the failure behaviour of weapon equipment on a military vessel, it may be useful to study the reliability of its composition. However, when insight is needed in for example the required yearly maintenance hours of a fleet, the individual properties and uncertainties of lower-level objects may need to be aggregated in order to ease the computational and analytical burden and thereby facilitate exploratory analysis on a higher level.

Now, although the combined use of models at multiple resolutions is potentially promising, few studies exist that work with models on different resolutions. In this light, it is currently quite difficult to state the merit of models at different resolutions. It may be true that a huge limitation of high-resolution models concerns their inability for exploratory analysis (Bankes, 1993). On the other hand, moving up the hierarchy of models inherently comes with information loss. Therefore, low-resolution models can never be considered superior to detailed models (Davis & Bigelow, 1998; Bigelow & Davis, 2003). In this light, this study aims to bring both worlds together by answering questions about (a) how models at higher resolutions can be aggregated, (b) how models at different resolutions can be compared, and (c) how models at different resolutions can be used to decrease each other's weaknesses and increase mutual strengths.

The aim of this study is to understand how multi-resolution modelling (defined by Davis & Bigelow (1998, pp. 5) as "*building an integrated family or two or more mutually consistent models of the same phenomena at different levels of resolution"*) can aid in improving the robustness of a force support system and thereby increase the performance of a force. As a first step, a force support model will be developed that serves as an aggregation of an existing lower level force support model developed for the Dutch Royal Navy. Second, it is checked how consistency between models can be achieved. Based on this cross-validation process, questions are answered regarding the comparability of models. Consequently, with EMA techniques, this study tries to (a) find scenarios on an aggregated level that lead to a low performance of a force, and (b) implement and evaluate policies on an aggregated level in order to close a potential performance gap. All in all, this study aims to answer the following question:

"How can multi-resolution modelling be useful in improving the robustness of a naval force support system?"

The main contribution of this study concerns the multi-resolution modelling (MRM) approach in researching the performance of a naval force support system. In contrast to previous research, this study is one of first studies with the ambition to improve the performance of a naval force support system by working with different levels of aggregation. This has implications for how currently existing force support models at different resolutions are used. Although this study does not aim to generalize findings to other cases using MRM, it is believed that valuable lessons can be learned from this study with respect to the aggregation, validation, and use of models at different resolutions. This study also has practical implications for the design and use of the force support model developed for the Dutch Royal Navy.

This research is structured as follows. The second chapter presents the required methods and subquestions in order to answer the main question of this study. A variety of methods are required for the aggregation and validation of models, but also for modelling & simulation practices and for exploratory analysis. The third chapter discusses the aggregation of a high-resolution model, and in particular the aggregation of heterogenous objects. The fourth chapter answers questions regarding the comparability of models by subjecting them to a variety of (cross-)validation tests. The fifth chapter provides an example case that illustrates the benefit of having an aggregated model. Chapter six provides an academic reflection about the use of MRM methods in this study and hints to further research directions. Chapter seven provides a conclusion of how a low-resolution model and a high-resolution model can be collectively useful in improving the robustness of a force support system. Chapter seven also includes the limitations of this study, and a number of scientific research recommendations. The practical implications of this study remain confidential and are not included in this report.

2. Methodology

The first four sections contain a description of (1) the MRM methods, (2) the approach to achieve a robust force support system, (3) the modelling and simulation methods, and (4) the case used in this study. Section 2.5. describes the research flow, including the subquestions.

2.1. Multi-Resolution Modelling

Three core concepts in the multi-resolution modelling (MRM) field are relevant for the aggregation and validation process: (a) the dimensions of resolution, (b) Integrated Hierarchical Variable Resolution (IHVR) modelling, and (c) the consistency framework. First, the dimensions of resolution (Davis & Hillestad, 1993) are used to comprehend the resolution of the detailed model used in this study and to consequently determine where aggregation is needed. In short, resolution is divided in four main dimensions: object-related, process, spatial, and temporal. The object resolution primarily refers to the granularity of the units under consideration (e.g. an installation, a vessel or a fleet) but also involves the detailedness of the dependencies between objects. Process resolution can refer to detailedness of the asset degradation process in a force support system. Finally, spatial and temporal resolution refers to fineness of the scale of space and time.

Second, this study adopts Integrated Hierarchical Variable Resolution modelling (IHVR) as an aggregation method. When feasible, IHVR is one of the most effective aggregation methods as it aims to achieve consistency between models in a highly systematic way (Davis & Hillestad, 1993; Davis & Bigelow, 1998). In short, the basic concept of IHVR is that variables are described in hierarchical trees with the highest-resolution variables at the bottom and the lowest-resolution variables on top. The multi-resolution variables are connected with mathematical aggregation functions which 'transform' higher-resolution variables to lower ones. Due to the use of intermediate aggregations, aggregation functions can be kept relatively simple which increases the transparency and interpretability of the aggregation process.

Regarding IHVR, this study distinguishes two types of aggregations: Case A aggregations and Case B aggregations (Davis & Bigelow, 1998). In Case A, the high-resolution model can be aggregated without a need for new approximations. In Case B, it is more difficult to rewrite the high-resolution structure to a lower resolution as consistency errors will be introduced. In fact, different approximations may be needed in the low-resolution model in order to represent the same phenomena in the high-resolution model. The errors due to such Case B aggregations cannot be 'too big' in order to keep models 'sufficiently' consistent.

Third, consistency between models at different resolutions is generally divided in weak consistency and strong consistency (Davis & Bigelow, 1998; Bigelow & Davis, 2003). Achieving strong consistency is not feasible within the available research time as it also involves disaggregation; it would however be an excellent step after this research. Weak consistency is achieved when one (a) aggregates the initial high-resolution state of a model to an initial aggregate state, (b) runs the high-resolution model and the low-resolution model, (c) aggregates the final high-resolution state to a final aggregate state, and (d) finds the error between the result of the aggregated run and the aggregated result of the high-resolution run to be 'sufficiently' low. Although model consistency is a well discussed topic in MRM literature, little 'practical' case studies exist that specify when models can be considered 'sufficiently' consistent (Davis, 2016). Therefore, this study aims to achieve consistency in a more systematic way by specifying cross-validation tests in section 2.5.

2.2. Exploratory Modelling & Analysis

In this study, the aggregated model is used for exploratory analysis. As already touched upon in the introduction, Exploratory Modelling & Analysis (EMA) is a methodology that uses computational experiments to analyse problems characterized by deep uncertainty (Bankes et al., 2013). Deep uncertainty represents situations where the future states of a system are unknown (either due to epistemic or ontic uncertainty) and/or decision-makers cannot agree upon them (Lempert et al., 2003; Maier et al., 2016). For this study, the EMA workbench version 2.0.8. (Kwakkel, 2017) is used; a toolkit for exploratory modelling and analysis that is useful for problem conceptualization, scenario generation, and vulnerability analysis. First, the EMA workbench conceptualizes problems by means of the XLRM framework (Lempert et al., 2003). Regarding XLRM, X represents the uncertainties, L denotes the policy levers, R denotes the simulation model and M represents the model outcomes.

Given the conceptualized problem, scenarios are generated. For each scenario, a value is sampled from the bandwidth of each uncertainty, which is uniformly distributed. The use of a uniform distribution is common in exploratory analysis, as the assignment of different probabilities to different intervals is not possible in case of deep uncertainty (Walker et al., 2013). Regarding the sampling procedure, Latin Hypercube Sampling is preferred as this sampling method systematically covers the entire bandwidth of uncertainties (Helton & Davis, 2003).

Given a base ensemble, scenario discovery is used for vulnerability analysis. Scenario discovery tries to find parameters bandwidths that cause a particular subspace of an outcome of interest (e.g. worst case outcomes) (Bryant & Lempert, 2010). Scenario discovery is done with the PRIM algorithm (Friedman & Fisher, 1999) and dimensional stacking (Suzuki et al., 2015). PRIM is a supervised classification algorithm that evaluates regions (called boxes) in the uncertainty space for their coverage, density and interpretability. The coverage is the number of cases of interest within the box divided by the total number of cases of interest in the entire space. The density is number of cases of interest within a box divided by all the cases in the same box. The interpretability is represented by the number of uncertainties that characterize the box. In practice, a trade-off should be made between the coverage, density and interpretability when choosing the best box (Bryant & Lempert, 2010). Dimensional stacking is a more visual technique for performing scenario discovery, but has essentially the same purpose as PRIM.

Finally, it is important that exploratory models come with a strategy. According to Fitzsimmons (2006), explorative models can increase ambiguity and thereby preconceptions of defence decision-makers (Betts, 1982) when a great diversity of possible futures is presented without clear implications. In this light, this study will evaluate the robustness of policies by following the Adaptive Robust Design (ARD) framework by Hamarat et al. (2013). In short, the ARD framework involves an iterative cycle of (a) the identification of regions of interest (e.g. worst-case outcomes), (b) the design of policies, and (c) the evaluation of policies. In this study, only a single iteration will be performed. A better policy advice would involve more iterations or the adoption of robust optimization frameworks (Bartholomew & Kwakkel, 2020), but for the time being the only purpose of the aggregated model is to illustrate how it can be valuable for the design of a robust force support system.

2.3. System Dynamics

Regarding the force support models, System Dynamics (SD) is used as a modelling & simulation method. SD is a continuous simulation tool which falls under the Differential Equation System Specification formalism (Forrester, 1961; Zeigler et al., 2018). In short, SD aims to provide a holistic view of a dynamically complex system of interest and consequently simulates the resulting dynamics over time (Sterman, 2000). With SD, the central focus is on the link between the structure of a model and the behaviour over time emerging out of this structure (Lane, 2000). With such insights, SD can facilitate the design of systems by for instance the implementation of control structures (Wolstenholme, 1990).

The use of SD in defence planning is not new; Clark & Pisani (1985) and Coyle (1996, 1999) advocated it decades ago, as System Dynamics is proven valuable in understanding project failures involving many interdependent systems. As a matter of fact, force support planning involves many interdependent subsystems which makes the availability of a fleet complex. As interdependencies between maintenance policies, workforce availability, and fleet operations are mentally difficult to grasp due to feedback-loops and delays, decision-makers often push the system in the wrong direction when trying to eliminate a performance gap (Fan et al., 2010; Bowers et al., 2017). However, System Dynamics can provide insight in such complexity by modelling and simulating the causal structure of such a system (Forrester, 1973; Sterman, 1988). In this light, System Dynamics has extensively been used for force support planning in the past (Coyle & Gardiner, 1991; Adamides et al., 2004; McLucas et al., 2006) and it is still becoming more popular (Moallemi et al., 2018; Elsawah et al., 2018; Turan et al., 2020; McLucas & Elsawah, 2020).

Next to cross-validation tests, some ‘ordinary’ validation tests which are typical to SD models (Senge & Forrester, 1980) are applied in this study. Regarding structural tests, the structure-verification, parameter-verification, and dimensional consistency test are applied to the low-resolution model. The structure-verification test and parameter-verification test are passed when a model does not object structural knowledge or parametric knowledge about the real system. Furthermore, the dimensional consistency test checks for conflicting dimensions of model variables. Regarding behavioural validation, the behaviour-reproduction, the boundary-adequacy, and the extreme condition test are applied to the low-resolution model. For the behaviour-reproduction and boundary-adequacy test, an ensemble of runs of the low-resolution model is analysed. The behaviour-reproduction test checks whether behaviour generated by the model corresponds to behaviour observed of the real system. The boundary-adequacy test checks whether the model includes all relevant structures to serve its purpose. Finally, the extreme-conditions test is performed to evaluate whether conditions can be found where the model breaks. Please read section 2.5. for how ordinary validation relates to cross-validation and why both types of validation are necessary to establish the fitness for purpose of models at different resolutions.

2.4. Case description

The case study concerns a high-fidelity System Dynamics model developed by Copernicos for consolidative use by the Dutch Royal Navy. This model, hereinafter referred to as the WSMD (Weapon Systems Management Dynamics) model, aims to mimic the availability dynamics of a military vessel over its lifetime, based on a pre-specified mission profile. The WSMD model is coupled to a data base, from which vessel and mission configurations can be made. In principle, the availability of every Dutch military vessel can be simulated for a given mission schedule. For this research, one particular vessel will be chosen. Based on the availability (and costs) of a vessel, decision-makers aim to reason about the naval force design (e.g. vessel acquisition) as well as the design of the force support system (e.g. maintenance plans).

The Dutch MoD stated in a recent interview that half of the naval force is not operational due to shortages in materiel and human resources (Trouw, 2021). As will be discussed in chapter 3, rebuilding a robust force support system requires (a) reasoning on a strategic level, (b) exploratory analysis to make choices amidst uncertainty. However, this type of decision-making cannot be supported by the WSMD model. In general, the WSMD model operates on a high resolution; many different missions exist and the availability of a vessel is modelled in great detail. In this light, a strong aim exists to explore the performance of vessels on an aggregated level.

Now, a conceptual overview of the WSMD model is provided in figure 2.1. In the WSMD model, the vessel availability is made up of five interdependent submodels: missions, condition, age, failures, and recovery. In reality, the WSMD model is more elaborate but these submodels contain the main dynamics and therefore form the scope of this research. The following five paragraphs provide a brief description of each submodel.

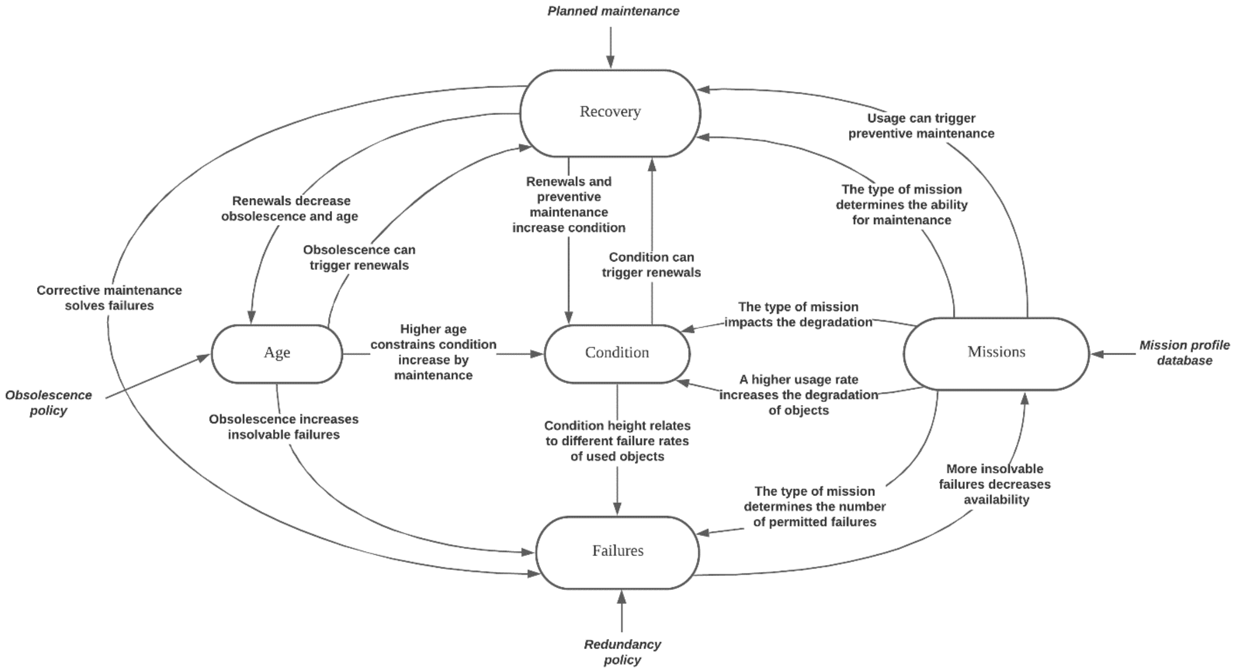
Planned maintenance
Recovery
Usage can trigger preventive maintenance
Renewals and preventive maintenance increase condition
The type of mission determines the ability for maintenance
Renewals decrease obsolescence and age
Obsolescence can trigger renewals
Condition can trigger renewals
Corrective maintenance solves failures
The type of mission impacts the degradation
Higher age constrains condition increase by maintenance
Age
Condition
Missions
Mission profile database
Obsolescence policy
A higher usage rate increases the degradation of objects
Obsolescence increases insolvable failures
Condition height relates to different failure rates of used objects
More insolvable failures decreases availability
The type of mission determines the number of permitted failures
Failures
Redundancy policy

Figure 2.1. Subsystem diagram (Morecroft, 1982) of the Dutch force support system in the WSMD model. Subsystems are shown in ovals, relations are represented by texted arrows, exogenous factors and policies are shown in italics.

The missions submodel provides input for the condition, maintenance, and failures of objects. Regarding the condition, each mission comes with a different impact on an object and a different required usage rate of an object (i.e. the percentage of time that an object is used during a certain mission). Both the impact and the usage rate influence the degradation rate. Furthermore, the type of mission determines the extent to which maintenance can be executed. To illustrate, a surveillance mission possibly allows the full maintenance potential of the crew, while maintenance during warfare missions usually amounts a fraction of the initial working capacity. Regarding the failures, every mission requires a different performance of objects and thereby a different number of 'allowed' unsolved failures. Finally, the extent to which objects are used determines the frequency of usage-based maintenance.

The condition submodel forms the core of the WSMD model. The condition decreases due to object degradation and can be increased with preventive maintenance or renewals. In case of preventive maintenance, the maximum achievable condition of objects is constrained by their age; a higher age lowers the upper condition boundary. Furthermore, objects require a renewal when their condition reaches a certain threshold. After that point, ordinary maintenance is deemed uneconomical or even technically impossible. A final important note is that the condition submodel is important for its self-reinforcing feedback effect; a lower condition leads to a higher degradation (Rashedi & Hegazy, 2016). This is an approximation of lower-level interactions between components; a lower condition of a component often affects the condition of others.

Unlike many standard reliability models (Dhillon, 2006), the WSMD model treats the Mean Time Between Failures (MTBF) endogenously by modelling the condition as the main driver of the failure generation. After failure generation, failures are solved with corrective maintenance. Each object has an in-built redundancy, meaning that it is immune to a certain amount of failures. When failures cannot be solved, either due to obsolescence or insufficient corrective maintenance capacity to meet the required performance, a ship becomes operationally unavailable.

Regarding the age submodel, an important concept is the obsolescence of objects. Obsolescence occurs when an object is no longer used even though it may still be in good working order (Hastings, 2015). The main reasons for obsolescence in the defence sector are (a) the inability to maintain or replace objects due to its dated structure, and (b) the disadvantage in efficiency or effectiveness of an object in comparison to the used technology by other (hostile) entities. In light of the second reason, the age submodel is quite exogenous as the technology development of other nations falls outside the scope of the WSMD model.

Finally, the recovery submodel represents the main interventions in the WSMD model: renewals and maintenance (the latter divided in preventive and corrective maintenance). Preventive maintenance can only increase the condition until the age constraint, while renewals also reset the age of the object. In principle, corrective maintenance has priority in case of failures; remaining maintenance hours (i.e. the working capacity minus the working hours for corrective maintenance) can be used for preventive maintenance. In general, preventive maintenance is quite exogenous in the WSMD model as it is mainly determined by static maintenance plans. Furthermore, each mission involves different maintenance crews which each have a unique capacity. A final important note is that the maintenance crew associated with the 'large maintenance' mission is the only crew that can renew objects.

2.5. Research flow

The question to which extent multi-resolution modelling is useful in improving the robustness of a naval force support system is divided in three questions, which will be described on the following pages. A schematic overview of the research flow is provided in figure 2.2.

1. *"How can the required aggregations on different dimensions of resolution in the WSMD model be implemented?"*

The answer to the first question provides information about the decision process behind model aggregation. For this question, it is important to explicitly state how model aggregations are implemented, as it forms the basis for how the WSMD model and the aggregated model compare. This question also involves the discussion of potential trade-offs in aggregation. Trade-offs usually relate to whether a loss of information is worth the aggregation in question, but also include computational cost considerations and interpretability. Please note that this question cannot be treated in isolation; the answer to the second question is decisive in deciding whether certain aggregations are possible or not. After applying the aggregation functions according to the IHVR method, the eventual product of this question is an implemented System Dynamics model that serves as an aggregation of the WSMD model.

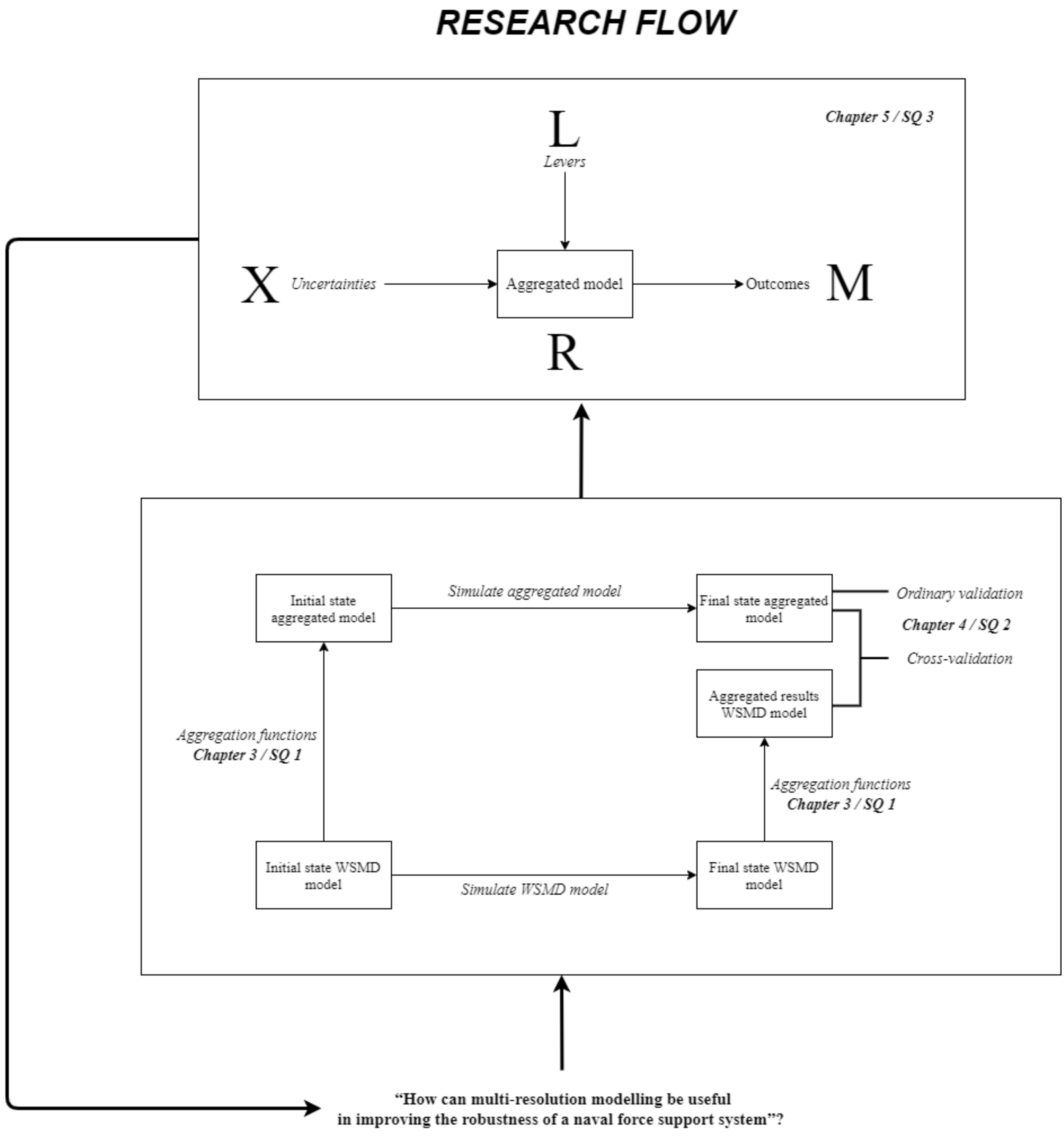

Figure 2.2. Research framework

2. *"How do the WSMD model and the aggregated model affect each other's validity?"*

This question aims to establish the validity of the aggregated model and the WSMD model by subjecting it to a variety of validation tests. Model validation is the process of building confidence in the usefulness of a model, which depends on its fitness for purpose (Senge & Forrester, 1980). In order to evaluate the fitness for purpose of the aggregated model, cross-validation tests are specified that compare the aggregated model to the WSMD model. Second, ordinary validation tests of Senge & Forrester (1980) are performed on the aggregated model only in order to identify new behaviours and insights that the WSMD model was unable to generate due to its limited ability for exploratory analysis.

Ideally, the initial parametrization of the discovered new behavioural modes in the aggregated model would be disaggregated in order to check whether the WSMD model can generate the same behaviour. However, due to the limited research time, this was infeasible and thereby remains a further research recommendation further described in section 6.1. Still, in the light of validation tests such as the boundary adequacy test, the validation of the aggregated model also has implications for the validity of the WSMD model. In this light, the validation phase is a two-way process which affects both the validity of the aggregated model as well as the WSMD model.

Regarding cross-validation, the main idea is that models are 'consistent'. In this research, models are considered consistent when they satisfy two conditions: (a) the aggregated model should not contradict the knowledge of the WSMD model regarding the modelled system of interest, and (b) the aggregated model should have a 'tolerable' numerical error with respect to the WSMD model. Note that condition (b) can only be satisfied after condition (a); therefore condition (a) serves as a prerequisite. Although a numerical error test is not common in SD models, the outcomes of aggregated model form a main indicator for the design of force support for the vessel in question. In this light, the numerical error between the WSMD model and the aggregated model cannot be too large. The specification of what error is tolerated is further discussed in chapter 4.

In order to satisfy condition (a), the behavioural modes identified in the WSMD model should not contradict the aggregated model. The fundamental behavioural modes presented by Yücel & Barlas (2011) will be used as reference material. These behavioural modes include different kinds of exponential behaviour, but also s-shapes and oscillations. Regarding the numerical error test, an important argument for the choice between different numerical error metrics concerns the capability of metrics to discriminate among model results (Chai & Draxler, 2014). For this research, it is deemed important that undesired deviations achieve higher weights in the error computation. In this light, the Root Mean Squared Error (RMSE) test is a solid choice as it discriminates high errors by means of squaring. Mathematically, the RMSE is expressed in this research as equation 2.1, where $WSMD_{o,t}$ and $AGG_{o,t}$ denote outcome *o* at time *t* for the WSMD model and the aggregated model respectively.

$$RMSE_o = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(WSMD_{o,t} - AGG_{o,t}\right)^2} \qquad (2.1)$$

3. "How can the aggregated model be used to increase the performance of the Dutch force support system as represented by the WSMD model?"

The purpose of this question is to illustrate the merit of the aggregated model with respect to the WSMD model. After stating the fitness for purpose of the aggregated model, experiments are performed in order to (a) get insight in worst-case scenarios regarding the vessel in question, and (b) evaluate policies in order to increase the robustness of the vessel support system. For this, the XLRM framework, the EMA workbench, and the ADR framework are used to conceptualize, generate scenarios, and find effective policies respectively. The robustness of policies on the outcomes are evaluated with robustness metrics, which are case-specific and therefore specified in chapter 5. Ideally, the policies on an aggregate level would be disaggregated and implemented in the WSMD model in order to study the effectiveness even better. Due to time limitations, the interchangeable use of high-resolution models and low-resolution models remains a point for further research (discussed in section 6.3.).

3. Aggregation

This chapter is structured as follows. First, the drawbacks of the WSMD model are highlighted with respect to its resolution. Second, after identifying the dimensions of resolution that are relevant for aggregation, the most 'compelling' aggregations are described in section 3.2. The third section contains a comprehensive description of the aggregated model. The fourth and final section contains the main message of this chapter in the light of the first subquestion.

3.1. Drawbacks of the WSMD model

From a resolution perspective, the WSMD model can best be categorized as a 'selective viewing' approach. With selective viewing, there is a single high resolution model from which aggregations to high level variables are made (Davis & Hillestad, 1993; Rabelo et al., 2015). Regarding the WSMD model, the high resolution is mainly characterized by the choice for individual installations (e.g. weapon equipment, navigation devices, propulsion engines etc.) on the object dimension and the choice for high-fidelity missions on the process dimension. Based on the performance of installations during a certain mission profile, aggregations are made to a higher level (e.g. the total maintenance costs of a vessel). However, although selective viewing allows 'zooming' and 'unzooming', Davis & Hillestad (1993) question whether selective viewing should be considered a multi-resolution modelling approach.

The main drawback of selective viewing is that even the highest aggregated results eventually depend on a high-resolution model. This dependency comes with many problems (Davis & Bigelow, 1998), but for the WSMD model two problems are particularly relevant. First, the WSMD model hinders the ability for higher level analysis (e.g. on a vessel level). To illustrate, a strategic question to investigate whether a 10% decrease in the sailing time also allows a 10% decrease in the maintenance hours of a ship is hard to answer as (a) the sailing time is composed of a variety of missions, and (b) installations have different maintenance plans which may overlap.

Second, the WSMD model does not facilitate exploratory analysis. In literature, the high computational complexity of detailed models is often presented as a main reason for this (Harrison et al., 2020). However, although computational power is a limiting factor for scenario generation with the WSMD model, the analytical costs are much more problematic. As communicated by the developers of the WSMD model, the collection and validation of input data poses a major challenge as in most cases the data is not available or it has a poor quality. Solving this issue by simply including each parameter in the uncertainty space heavily complicates the analysis of the output as the numerous inputs make cases much harder to describe and comprehend.

To summarize the previous two paragraphs, the use of a single low-resolution model is crucial as it (a) allows reasoning on a higher level, and (b) facilitates decision-makers to make choices amidst uncertainty. In this light, three dimensions of resolution formed the focus during the aggregation process: the object dimension, the process dimension, and the time dimension. Aggregations on these dimensions would heavily decrease the resolution of the WSMD model and thereby increase the ability for strategic analysis and exploration.

Before going into detail, the intended aggregations are shortly covered. Regarding the object dimension, installations are aggregated to a vessel level. As the aggregations on the object dimension involve case B aggregations, they are separately described in section 3.2. Regarding the process dimension, a number of aggregations are made. The most important aggregation concerns the aggregation of individual missions to generic operations. Individual missions facilitate much detail in the degradation, failures, and maintenance part of the WSMD model and therefore require aggregation. The process aggregations are covered in section 3.3, together with the description of the aggregated model. Finally, to ease the computational burden, the time dimension is aggregated from days to years. This choice on the time dimension does not require further explanation as strategic defence models are usually formulated in years (Gray, 2008).

3.2. Towards an aggregated model

As aggregation eminently comes with information loss, it is a major challenge to decide which high-resolution details can be omitted and which critical information should be kept (Bigelow & Davis, 2003). The main challenge of this section is to somehow capture the most relevant information of the installation mix in the WSMD model. Please note that every aggregation choice is checked with cross-validation tests in chapter 4. In other words, the applied aggregation functions are 'allowed' by the cross-validation tests with respect to their consistency error. Now, before moving on to the aggregations on the object dimension of the WSMD model, some notes are made about the process of finding and evaluating aggregation functions as this actually took most of the research time.

- Obviously, when working with aggregations, one should understand the high-resolution model. However, mentally grasping the WSMD model required many effort due to its detailedness and limited ability for computational experiments. In this light, many discussions were needed with the developers of the WSMD model.
- Finding appropriate aggregations was a trial & error process and required a combination of creativity, discussions with modelling experts, and adequate modelling skills.
- The evaluation of aggregation functions was time consuming. In particular, it took a lot of time to (a) compute the initial low-resolution state, (b) simulate the WSMD model, and (c) apply the cross-validation tests. Eventually, point (a) and (c) were automated with Python scripts which saved much time in the end.
- To continue on the previous point, a comprehensive evaluation of aggregation functions by cross-validating them on many high-resolution 'input regions' was undoable. The main reason for this was that the WSMD model and especially the coupled database did not allow for exploratory analysis. However, section 6.1 proposes a change in the consistency framework used in this study in order to avoid this problem in further research.

All in all, the entire aggregation process was deemed challenging. Now, section 3.2.1 conceptually describes the main aggregations on the object dimension, while section 3.2.2 covers their implementation.

3.2.1. Conceptual overview object aggregation

Although the main practical focus of this study is to aggregate the WSMD model, an earlier made aggregation on the object dimension (i.e. one level lower than the WSMD model) needs to be revised in order to prevent further consistency problems. To start with, the WSMD model formulates the attributes of installations in terms of averages of attributes of so-called 'component classes' (where each component class is an aggregation of individual components on a lower level). However, as each component class comes with its own unique behaviour, the conversion of attributes of component classes to a single attribute value on an installation level leads to a poor approximation. This can be supported with the cross-validation test of Yücel & Barlas (2011); when the behaviour patterns of component classes (especially with respect to their failure behaviour) are qualitatively different, an aggregation to a single behaviour pattern would definitely lose important dynamics. To illustrate, electronic components behave differently than mechanical ones by having less, or at least heavily delayed, wear out (or end-of-life) symptoms (Carchia, 1999).

In this light, it is decided to express installations in terms of component classes instead of treating installations holistically. This 'trick' to translate heterogenous objects to functionally similar entities (i.e. installations share the same component classes) is quite common in MRM. In a combat context for instance, Davis (1995) aggregates heterogeneous objects such as aircraft and tanks to 'shooters'. These types of aggregations can be at first sight confusing as the level of abstraction increases. However, according to Davis & Bigelow, *"normal object-oriented programming may actually be a hindrance to MRM because such normal practice — in which objects are identified with physical named systems — fails to exploit simplifying abstractions"* (Davis & Bigelow, 1998, pp. 43).

Now that this previous aggregation by the WSMD model is corrected, let us return to the object aggregation of an installation level to a vessel level. On a vessel level, the installation set is divided in two groups: installations that are prone to operational obsolescence and installations that are immune to operational obsolescence. To recap from section 2.4, operational obsolescence occurs when an object can no longer used even though it may still be in good working order (Hastings, 2015). To motivate the aggregation, installations that are prone to operational obsolescence require a renewal and therefore demand a different treatment during large maintenance than installations that are immune to operational obsolescence. As this results in diverging behaviour with respect to the condition and failures of installations, it is important to separate the 'obsolete' group from the rest. All in all, the object aggregations on a component level and an installation level can be summarized with figure 3.1. In the aggregated model, the objects aggregations result in m x n installation groups: dimension m represents the obsolescence attribute and dimension n represents the component classes.

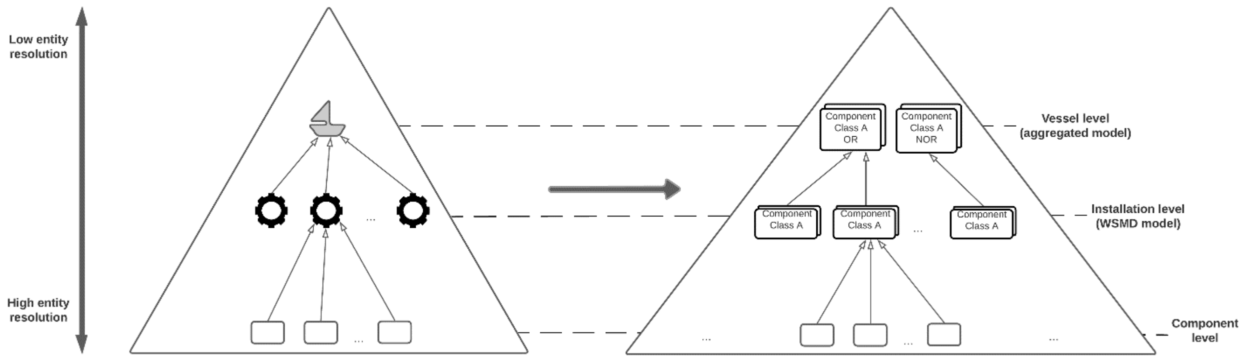

Figure 3.1. Visualization of the hierarchical composition of a vessel. On the left, a physical representation is shown. On the right, physical objects are represented from a multi-resolution modelling perspective. Regarding section 3.2.1, the representation on the installation level is discussed in the first two paragraphs while the representation on the vessel level is discussed in the third paragraph. On a vessel level, OR and NOR represent the installation groups that are prone and immune to operational obsolescence respectively.

3.2.2. Implementation object aggregations

Now, how should installations be aggregated to 'installation groups' on a vessel level? For some object aggregations, no new approximations (i.e. Case A aggregations) were required on an aggregate level. This was the case when installations had similar attribute values or when their uncertainty bandwidths largely overlapped. This was for instance the case for the obsolescence free period, an array that denoted the time of no obsolescence for each installation. Furthermore, no new approximations were required when heterogenous parameters in the WSMD model had a linear impact on the cross-validated outcomes. That is: *E(f(x)) = f(E(x))*. In such cases, simple aggregation functions such as the mean could be implemented. In other cases however, such aggregations did not hold due to non-linearity in the WSMD model. In such cases, case B aggregations were required.

Before moving on to the implementation of case B aggregations, some background about the current structure of the WSMD model is required. To start with, the WSMD model aims to represent attributes (e.g. the condition, age etc.) of objects with a comprehensive stock. Following this line of reasoning, the use of a comprehensive stock on a vessel level implies that attributes of installations are aggregated to a single attribute of a vessel. However, due to the heterogeneity of installations, the use of a comprehensive stock appeared to be problematic on a vessel level. Moreover, the use of a comprehensive stock in the WSMD model on an installation level is questionable as well, as higher object resolutions than installations also involve heterogeneity. The following paragraphs aim to illustrate the problem of using a comprehensive condition stock on different levels of aggregation. Please note figure 3.2 first.

Considering figure 3.2, suppose that the black line in the graph on the left represents the condition course of an installation. At some moment in time, an intervention takes place and the conditions of individual components are increased. Of all components, 70% gets ordinary maintenance (e.g. greasing) and 30% needs to be renewed. However, as the condition of the component class is modelled as a comprehensive stock in the WSMD model, there is no other possibility than taking an equivalent condition increase (e.g. by computing the average increase) for all individual components. This is problematic as the condition relates to different phases in the failure behaviour of components.

In general, asset management literature distinguishes three different phases in the failure behaviour of assets: (1) the infant phase where the failure rate usually exhibits a negative exponential decline, (2) the normal phase which is usually characterized by a constant failure rate, and (3) the wear out phase which usually exhibits a failure rate with a positive exponential growth. Together, the phases make up the famous 'bathtub' model (Dhillon, 2006; Hastings, 2015). Now, returning to the example, 30% of the components is renewed and therefore arrive in the infant phase while the rest returns to the normal phase. However, the average approximation in (B) will entirely result in a failure behaviour that belongs to the normal phase. Therefore, the approximations (A) and (B) result in different failure rates; both with respect to the measured numeric difference between the failures and the difference in behavioural patterns.

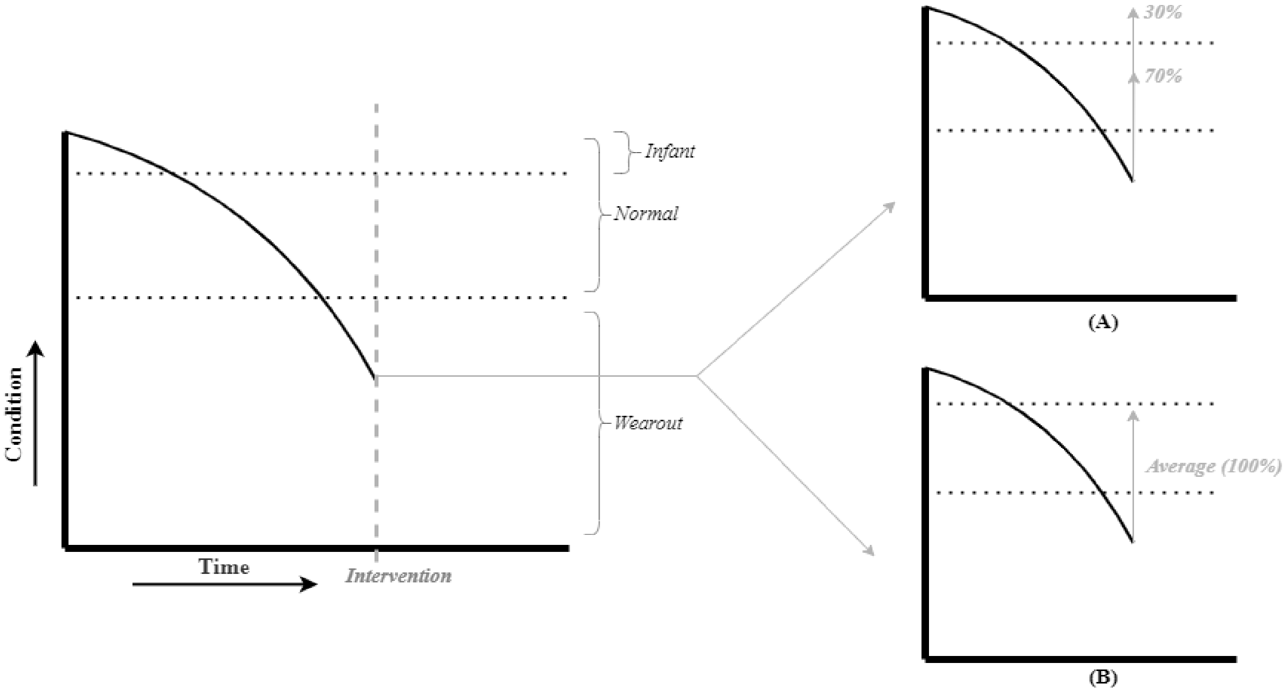

Figure 3.2. Illustration of a potentially problematic approximation (B) and a more appropriate approximation (A) of the situation presented in the left graph. The intervention is denoted with a vertical grey dotted line. The horizontal dotted lines represent different phases in the failure behaviour of a component class. Please note that the representation of phases is a simplification; in reality phases are way less discrete and not every phase is (only) dependent on the condition.

The continue, a similar problem occurs when aggregating the condition of individual installations to the condition of a vessel. Due to a variety of factors, installations have a different degradation rate and thereby a different condition. When using a comprehensive stock, a very slowly degrading installation and a highly deteriorating one result in a different failure rate than the failure rate that is mapped from the condition on a vessel level. Of course, in reality the failure phases will not be as clear-cut as described, but the main point remains that glossing over qualitatively different phases results in poor approximations. Fortunately, a combination of 'abstract' aggregations and the use of different modelling structures may solve the in-group heterogeneity problem. Note that this aggregation is only performed from an installation level to a vessel level, and not of a component level to an installation level. The latter would require cross-validation with the WSMD model and a model with a higher object resolution than the WSMD model, but this is not within the scope of this research.

In the aggregated model, a variety of stocks is implemented that represent different parts of the attribute range of objects. In System Dynamics literature, such a concept is also known as a 'chain' and it is often used for population dynamics (Auping et al., 2015). Regarding the condition, the condition range is divided in a number of 'stages' which each represent a different phase in the failure generation of objects. With the use of an 'average' degradation rate, installation groups on a vessel level are distributed over the set of condition stocks, thereby representing the heterogeneity in the degradation of installations on a lower level. To make this more concrete, section 3.3.2 provides insight in the formal aggregation process of the chain approximation.

However, the use of such a chain only did prove to be insufficient during the cross-validation phase. In order to capture the heterogeneity better, the attributes of installations are first converted to component classes before being aggregated to a vessel level. To illustrate, instead of a single usage rate, a number of usage rates (i.e. one for each component class) existed for an installation group. As described in section 4.1.2, the numeric error between the aggregated model and the WSMD model remained sufficiently low due to this aggregation. However, such aggregations are not always possible. In case that low-level heterogeneity cannot be captured by creating more groups on a higher level, the validity of the aggregated model changes. This is further discussed in section 4.2. A case where the creation of additional groups was infeasible concerned the obsolescence structure in the WSMD model, which will be discussed in section 3.3.3.

Regarding the use of a chain and abstract aggregations in general, a few trade-offs should be mentioned. First, the use of a chain requires trade-offs in the computational speed and the accuracy (i.e. the consistency error) of results. As an one-to-one representation of the entire condition range in the WSMD model would require an infinite amount of condition stages in the aggregated model (where the number of stages proportionally increase the running time), it should be noted that a chain eminently comes with errors in approximating the condition and failures of objects. The choice for the amount of stocks then depends on the notion which error is acceptable with respect to a comprehensive stock. Chapter 4 further elaborates on the comparability between the chain in the aggregated model and the use of a comprehensive stock in the WSMD model.

Furthermore, abstract aggregations decrease the interpretability of the aggregated model. This may be problematic when the aim of the aggregated model was to increase the transparency of the higher-resolution models. For this case, decision-makers cannot specify an average usage rate for an installation group, but have to specify an average usage rate for each component class of each installation group. This requires an extra step 'mental' step, as installations first need to be mentally translated to their composition. Having too much 'mental' steps on an aggregate level may impede the interpretability, which is actually a huge strength of low-resolution models (Davis & Bigelow, 1998). Therefore, such aggregations should always be implemented in agreement with the intended users of the aggregated model in question.

3.3. Aggregated model description

The subsystem diagram in figure 3.3 displays the dynamics of the aggregated model, as well as the omitted and additional information compared to the WSMD model. As the WSMD model is primarily based on settled theory in the asset management field and reliability engineering field, most implementations in the aggregated model do not need scientific back-up. Note that information described in section 2.4 will in principle not be repeated unless some structure changes in the aggregated model. The aggregated model contains around 150 equations and 44 subscripts, and is implemented in Vensim DSS version 8.1.0 (Ventana Systems, 2010). Before describing the model, a short note about the aggregation functions on the process dimension.

On the process dimension, most aggregations had to deal with heterogeneity in missions (as missions were aggregated to operations, described in section 3.3.1). However, during cross-validation, many aggregation functions could not or poorly be evaluated due to (a) the limited ability of the WSMD model to do multiple runs with different initialisations (as the specification of the inputs in the database is time intensive and the WSMD model also has high running time), and (b) on-going developments in the failure and maintenance submodel of the WSMD model, which hindered the inclusion of the recovery submodel in the cross-validation process. Nonetheless, the cross-validation functions are 'hypothetically' applied during 'structural' cross-validation, further described in section 4.2.

3.3.1 Operations submodel

The mission submodel in the WSMD model is decoupled from its static database and the schedule is aggregated to three types of operations: sailing, intermediate maintenance and overhaul (or large maintenance). This choice is motivated by the fact that these operations are associated with different dynamics. To illustrate, maintenance during sailing missions and intermediate maintenance does only involve corrective and preventive maintenance while renewals can only take place during overhaul. Also, each operation is associated with a different crew which in turn have a unique working capacity. All in all, these distinctions make each operation unique and thereby unsuitable for further aggregation.

The implementation of a schedule is heavily simplified in the aggregated model. In the WSMD model, the mission schedule is hardcoded in a database and forms a dynamic input to the model parameters. Mathematically, the mission schedule in the WSMD database can be represented as a binary *t x m* matrix **MS** with dimension *t* as the simulation time and dimension *m* as the mission types. Element $a_{t,m}$ denotes whether mission *m* is executed on time *t*. A formal notation is given in equation 3.1.

$$\mathbf{MS} = \begin{bmatrix} a_{t=0,m=1} & \cdots & a_{t=0,m=m} \\ \vdots & \ddots & \vdots \\ a_{t=t,m=1} & \cdots & a_{t=t,m=m} \end{bmatrix} \tag{3.1}$$

Needless to say, the parametrization of new mission schedules is time intensive due to the countless dynamic inputs which inhibit the ability for exploratory analysis with the WSMD model.

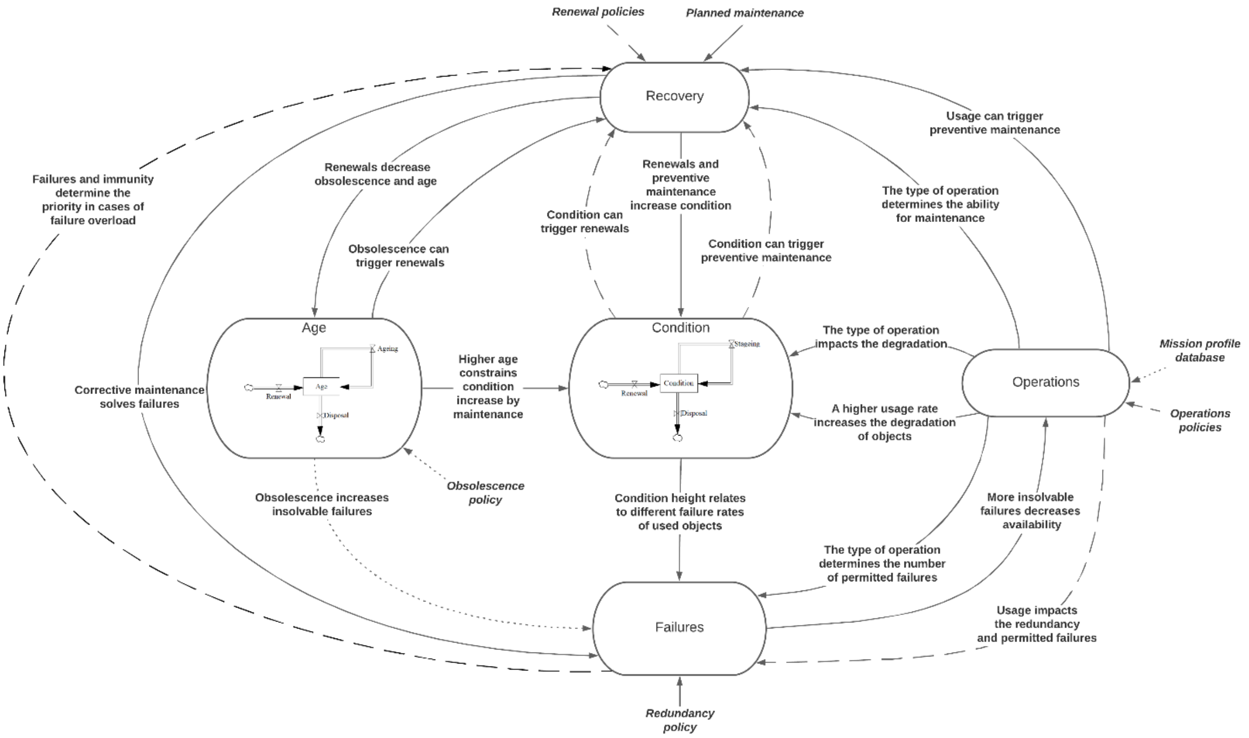
Renewal policies
Planned maintenance
Recovery
Usage can trigger preventive maintenance
Failures and immunity determine the priority in cases of failure overload
Renewals decrease obsolescence and age
Renewals and preventive maintenance increase condition
Condition can trigger renewals
The type of operation determines the ability for maintenance
Obsolescence can trigger renewals
Condition can trigger preventive maintenance
Age
Ageing
Age
Renewal
Disposal
Higher age constrains condition increase by maintenance
Condition
Staging
Condition
Renewal
Disposal
The type of operation impacts the degradation
Mission profile database
Operations
Corrective maintenance solves failures
A higher usage rate increases the degradation of objects
Operations policies
Obsolescence policy
Obsolescence increases insolvable failures
Condition height relates to different failure rates of used objects
More insolvable failures decreases availability
The type of operation determines the number of permitted failures
Failures
Usage impacts the redundancy and permitted failures
Redundancy policy

Figure 3.3. Subsystem diagram (Morecroft, 1982) of the aggregated model of the WSMD model. Subsystems are shown in ovals, relations are represented by texted arrows, exogenous factors and policies are shown in italics. With respect to the WSMD model, additional relations are indicated with a long dash and omitted relations are indicated with a short dash.

Therefore, in contrast to the WSMD model, the ‘new’ operations submodel is more suitable for exploratory analysis as an operation schedule can be developed with the use of only four parameters:

- The duration of an overhaul operation
- The cumulative time between two consequent overhauls (hereinafter referred to as the overhaul interval)
- The cumulative intermediate maintenance time in an overhaul interval
- The number of stops at a harbour for intermediate maintenance in an overhaul interval

Given the WSMD database, the parameters can be derived with database operations in Python. It should be noted that this simplification assumes that an operation has a fixed (average) duration. However, although missions can have slightly varying lengths in practice (and therefore in the WSMD database), this information loss appeared to be insignificant after applying the cross-validation tests (as far as the aggregation functions could be evaluated).

A final note is that the usage rate of installations becomes important when determining the redundancy and required performance on a vessel level. If the usage rate is not included, unused installations which do not cause failures mistakenly increase the total redundancy or required performance of a ship. Therefore, the usage rate is included in the aggregation of the redundancy and required performance of installations in the WSMD model.

3.3.2. Condition, age (part 1), and maintenance (part 1) submodels

The structure of the condition submodel and age submodel is changed from a comprehensive stock representation in the WSMD model to a chain representation in the aggregated model. In the WSMD model, the comprehensive stock modelled the condition and age of an installation. In the aggregated model, a vessel has multiple conditions and ages as multiple stocks exist (that all may contain a share of the vessel). In the aggregated model the condition chain and age chain are integrated, resulting in a two-dimensional chain. First, the age chain consists of twelve different stocks (called ages), where the first five ages represent three-year cohorts and the final seven ages represent five-year cohorts. The ageing time of each age stock is equal to its associated cohort value (i.e. 3 years or 5 years). The condition limitation due to age of each age stock is equal to the cumulative average value of each cohort (e.g. for three-year cohort 3 this value equals 10.5 year) multiplied with the output of a graphical function that maps the relation of the age to the associated condition decrease.

Second, the condition chain consists of twenty different stocks (called stages) that each cover an equal share of the total condition range. The condition chain also involves a calibration vector, but this covered in section 4.1.2. The degradation and recovery of objects imply that objects can flow to both lower stages and upper stages. Similar to the WSMD model, the aggregated model implements a reinforcing feedback-loop between the condition and the degradation rate. The evaluation of this feedback-loop with respect to lower-level interactions (represented in models with a higher resolution than the WSMD model) would be interesting, but this is not within the scope of this study. Furthermore, each stage in the chain has an associated condition value. The associated condition value of each stage differs per approximation. For the computation of the recovery time and the degradation time of each stage, the average stage condition value is taken. For the failure mapping, the associated condition values to each condition stage are different; please read section 3.3.5.

As the aggregation of a comprehensive condition stock to a condition chain remains abstract, this paragraph aims to further clarify this. To simplify a bit, assume that the condition is only determined by the degradation of installations. In the WSMD model, the condition of installations is then given by equation 3.2.

$$\frac{dc_i(t)}{dt} = -deg_i\,(t) \qquad where\ c_i(0) = 1 \tag{3.2}$$

Where $c_i(t)$ and $deg_i\,(t)$ are the condition and degradation rate of installation i respectively. The condition is quantified in 'condition unit' (cu) and the degradation rate in cu/day. A simplified representation of the degradation rate of an installation is given by equation 3.3.

$$deg_i(t) = \left(C_1 - c_i(t)\right) * C_{2,i} \qquad where\ C_1 > c_i(0) \tag{3.3}$$

Where C_1 (unit: cu) and $C_{2,i}$ (unit: 1/day) are for now assumed to be degradation constants (in practice, the degradation rate depends on both static and dynamic factors and it also switches to 0 at some time instant as the condition cannot be negative). Now, a simplified representation of the aggregation of the condition in the WSMD model is visualized in figure 3.4.

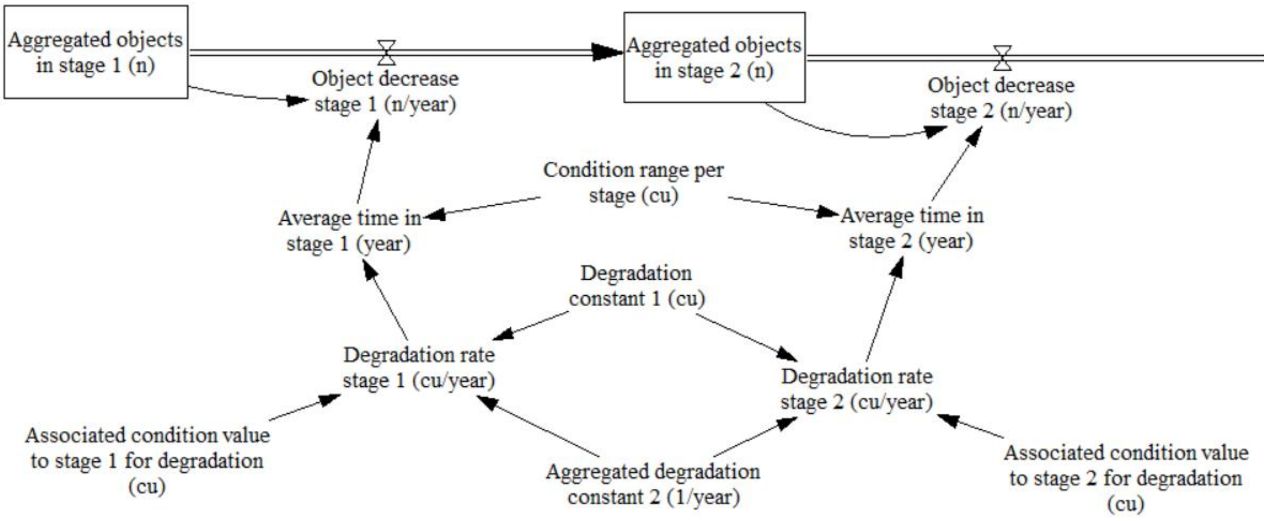
Aggregated objects in stage 1 (n)
Object decrease stage 1 (n/year)
Aggregated objects in stage 2 (n)
Object decrease stage 2 (n/year)
Condition range per stage (cu)
Average time in stage 1 (year)
Average time in stage 2 (year)
Degradation constant 1 (cu)
Degradation rate stage 1 (cu/year)
Degradation rate stage 2 (cu/year)
Associated condition value to stage 1 for degradation (cu)
Aggregated degradation constant 2 (1/year)
Associated condition value to stage 2 for degradation (cu)

Figure 3.4. Simplified representation of the condition chain in the aggregated model. Only the first two stages (of twenty in total) are shown. Units are shown in brackets.

In the aggregated model, the comprehensive condition stock is 'cut' in twenty stages where each stage covers a condition range of 0.05. As each stage stock integrates the object population, equation 3.2 changes to equation 3.4 in the aggregated model.

$$\frac{dpop_{s,ig}(t)}{dt} = \frac{pop_{s,ig}\ (t)}{\bar{t}_{s,ig}} \tag{3.4}$$

Where $pop_{s,ig}(t)$ and $\bar{t}_{s,ig}$ are the aggregated population and average time of installation group ig in stage s respectively. The average time in a stage is given by equation 3.5.

$$\bar{t}_{s,ig} = \frac{CR_s}{deg_{s,ig}} \tag{3.5}$$

Where CR_s is the stage condition range (in this case each stage has an equal range of 0.05) and $deg_{s,ig}$ is the aggregated degradation rate. The aggregated degradation rate of each stage is computed according to the same form of equation 3.3. However, in the aggregated model the degradation constant C_2 is an aggregated parameter, and condition $c_i(t)$ is replaced with a fixed condition value C_s for each stage. To illustrate the latter, the associated condition values of the first two stages are 0.975 and 0.925 respectively for the degradation mapping (i.e. the average of the covered condition range of stage 1 [1-0.95] and stage 2 [0.95-0.90]). All in all, in the aggregated model equation 3.3 changes to equation 3.6.

$$\begin{gathered} deg_{s,ig} = (C_1 - C_s\ )\ *\ C_{2,ig} \\ where\ C_1 > C_{stage1} \end{gathered} \tag{3.6}$$

Comparing equation 3.3 with equation 3.6, one should note that consistency errors are introduced as C_s is an approximation of $c_i(t)$. Chapter 4 further elaborates on the error between using a comprehensive stock versus a chain.

To continue on other aggregations, the external degradation part of the WSMD model is aggregated to a single exploratory parameter. In the WSMD model, multiple external impact types existed. However, the existence of significant overlapping uncertainties around each impact type stressed the cognitive need for fewer parameters. Next, regarding the recovery part, the recovery during sailing operations and intermediate maintenance is separated from the recovery during overhaul. During overhaul, objects are either renewed or recovered until their maximum achievable condition. Note that due to the chain in the aggregated model, partial object renewals can also take place (this is not possible in the WSMD model and therefore this relation is dashed). During sailing or intermediate maintenance, only the available amount of preventive maintenance can be executed (i.e. the part that remains after corrective maintenance).

3.3.3. Age submodel (part 2)

Regarding the obsolescence structure, some information is lost in the aggregation process. In the WSMD model, multiple types of object obsolescence exist while only operational obsolescence is included in the aggregated model (recap section 3.2.1). The reason for this was that including all obsolescence types would result in too many installation groups. As an installation can be prone to multiple obsolescence types, creating exclusive installations groups would approximately result in the same number of groups on a vessel level as the initial number of installations. In that case, aggregation becomes pointless. Therefore, the aggregated model only includes operational obsolescence, as this is the only obsolescence type that has an indirect relation to the condition and age of an object. When an installation becomes operationally obsolete, it requires a full replacement which results in a reset of its age and condition.

However, this choice has consequences for the structure of the aggregated model. In the WSMD model, obsolete installations can increase the number of insolvable failures. This holds for all obsolescence types, except for operational obsolescence. In this light, the relation between the age submodel and the failures submodel disappears. The same holds for the obsolescence policy; a policy on an installation level that could delay the obsolescence intervention moment of each individual installation for all obsolescence types except operational obsolescence. For an elaborate description of the consequences of this information loss, please read section 4.2.

3.3.4. Recovery submodel (part 2)

Regarding the rest of the recovery submodel, two changes are made in aggregated model with respect to the WSMD model. First, an adaptive maintenance type is introduced: condition-based maintenance (CBM). In the WSMD model, the initial plan was to include CBM as well, but this was effective nor efficient as it was executed at the wrong time instants as it was unclear when the static maintenance plans were (fully) executed. In the aggregated model however, CBM directly depends on the condition gap. The condition gap is specified as the aimed condition of CBM minus the condition of the installation groups. Ideally, the aimed condition of CBM should be based on some optimization procedure, but this remains a point of further research. Furthermore, CBM comes with some additional (fixed) manhours for the monitoring of the condition. In the aggregated model, CBM is modelled as an alternative for the initial maintenance plans with the use of a switch. In this light, the effectiveness and efficiency of the current maintenance plans can easily be evaluated.

Third, the aggregated model uses a different way in which corrective maintenance and preventive maintenance is divided over the component classes in case that the required maintenance exceeds the available capacity. In the WSMD model, each installation received an equal share in the available maintenance capacity in such cases. In the aggregated model, the maintenance share of installation groups is based on the maintenance priority. For corrective maintenance, installation groups have priority when they have many failures and a low immunity. For preventive maintenance, installation groups have priority when they have the largest relative (i.e. independent of the volume) share in the required preventive maintenance stock. Note that the latter change is not shown in figure 4, as this feedback-loop is within the recovery submodel. These changes in the aggregated model have no implications for the cross-validation process as they will also be implemented in the WSMD model.

3.3.5. Failures submodel

In the aggregated model, the associated condition values of the stages are mapped to a failure rate. For stages that cover condition values above 0.5, the upper boundary is taken as an associated value, while for stages that cover condition values below 0.5, the lower boundary is taken. For example, if stage 10 and stage 11 cover condition ranges 0.55-0.5 and 0.5-0.45, then the associated condition values denote 0.55 and 0.45 respectively. For the mapping, graphical functions are used that represent the effect of the condition on the nominal MTBF of component classes. The graphical functions only contain the normal phase, which starts at the highest possible condition value, and the wear-out phase, which ends at the lowest possible condition value (for a generic description of the failure phases, please recap section 3.2.2). The graphical functions are developed with expertise of the developers of the WSMD model and with literature in the field of reliability engineering (Dhillon, 2006; Gorjian et al., 2010; Karandaev et al., 2014; Carchia, 2019).

Regarding the infant phase, a separate stock is modelled that increases in case of renewals and decreases according to the average time in the infant phase of component classes. The infant phase is separated from the other phases as the duration of the infant phase does generally not depend much on the condition but rather on the object design and the tech familiarity by the crew (Adamides et al., 2004; Ahram et al., 2017). All in all, the eventual failure rate on a vessel level is equal to the sum of the condition stages multiplied with the associated failure rate, plus the additional infant failures (and then, of course, summed over all installation groups).

3.3.6. Key Performance Indicators (KPI's)

With an eye upon the robustness analysis in chapter 5, the aggregated model specifies two main KPI's: the availability of the vessel and the working hours. First the vessel availability is specified as the realized sailing time divided by the planned sailing time. Second, the working hours is divided in a maintenance part and a renewal part. The maintenance hours and renewal hours are separately summed over all different maintenance crews. Although both outcomes have the same units, they cannot be summed as they involve different costs (renewals are in general more expensive than maintenance activities). As the cost structure in the WSMD model does not fall under the scope of the aggregated model, these outcomes are separated. The availability and working hours outcomes are expressed in equation 3.7 and equation 3.8 respectively. Regarding the working hours, E denotes the set of maintenance crews e, and t denotes either renewal hours or maintenance hours.

$$sailing\ availability = \frac{realized\ sailing\ time}{planned\ sailing\ time} \tag{3.7}$$

$$working\ hours_t = \sum_{e=1}^{E} working\ hours_{t,e} \tag{3.8}$$

3.4. Key takeaways of aggregation

- The development of an aggregated model next to the WSMD model is deemed important as it (a) allows reasoning on a higher level, and (b) facilitates decision-makers to make choices amidst uncertainty by means of exploratory analysis.

- Important aggregations that simplified the WSMD model concerned the object dimension and the process dimension. On the object dimension, installations are aggregated to installation groups. On the process dimension, real-world missions are aggregated to three generic operations: sailing, intermediate maintenance, and overhaul (i.e. large maintenance).

- Aggregations on the object dimension required new approximations in the aggregated model. More concrete, assigning installation groups an equivalent age and condition does lead to unacceptable consistency errors. In order to better capture non-linearity and installation heterogeneity, the following aggregation operations are performed.

    1. Installations groups are converted to component classes. Component classes can better capture the lower-level heterogeneity as they represent the composition of installations in the installation group.
    2. Instead of using a comprehensive stock for the age and the condition, the aggregated model implements a 'chain' of condition stages and age cohorts that allows lower-level objects to have different attribute values.

- Implementing 'abstract' groups on an aggregate level to capture object heterogeneity on a lower level can decrease the consistency error but impede the interpretability of a low-resolution model.

- Implementing a detailed chain on an aggregate level to capture non-linearity and object heterogeneity on a lower level can decrease the consistency error but increase the computational complexity of a low-resolution model.

4. Validation

Model validation is the process of building confidence in the usefulness of a model (Senge & Forrester, 1980). The usefulness of a model depends on its fitness for purpose, which does not necessarily imply a high similarity between the model data and the real-world data. This chapter aims to evaluate the fitness for purpose of the aggregated model by subjecting it to a variety of validation tests. However, without stating the purpose first, it is pointless to establish the merit of a model or to compare the merits of different models (Senge & Forrester, 1980; Bigelow & Davis, 2003). In this light, the following paragraphs aim to state the intended purpose of the WSMD model and the aggregated model.

The main purpose of the WSMD model is to investigate whether the maintenance plans align with the intended usage of a vessel in order to realize the planned sailing missions. The type of decision-making associated with the WSMD model is tactical in nature. Operational decision-making would concern an accurate representation of the availability during a single mission, while strategic decision-making would concern generic operations rather than specific missions. As communicated by the WSMD model developers, the associated timeframe of the WSMD model is in months.

In order to serve its purpose, the WSMD model is based on settled theory of the reliability engineering field and a range of (primarily parametric) assumptions. Furthermore, the WSMD model has been empirically tested against real-world mission data (failure data in particular). However, this data only captures a very limited range of the possible applications of vessels at hand. Also, as conditions change over time, the data used for validation in the past may not be the same for future missions. All in all, taking into account the limited ability for exploratory analysis, it is the question whether the WSMD model fits its purpose.

In this light, the aggregated model aims to extend the usefulness of the WSMD model by enabling the exploration of a wide range of scenarios. Furthermore, as discussed in section 3.1.1, a clear need exists to develop a model that facilitates strategic decision-making. The need for exploratory analysis and strategic reasoning required the aggregations stated in the previous chapter. However, as aggregations lead to less complexity and therefore a loss of information (Bigelow & Davis, 2003), it is unclear how the aggregated model compares to the WSMD model. Eventually, it may turn out that for some cases the aggregated model is not comparable with the WSMD model.

In this study, cross-validation tests provide insight in how the aggregated model differs from the WSMD model, and whether this affects the fitness for purpose of the aggregated model. The aggregated model is cross-validated on two internal variables: the condition and the failure rate. These variables make up the core of the WSMD model. It would be better to cross-validate the models on their KPI's, but this is currently not possible. Due to on-going developments in the failure and maintenance mechanisms in the WSMD model, the recovery submodel and a large part of the failure submodel cannot be formally included in the cross-validation process which hinders cross-validation on the KPI's.

In this light, a distinction is made between behavioural cross-validation in section 4.1. and structural cross-validation in section 4.2. Behavioural cross-validation expresses the comparability of models with the use of cross-validation tests. In contrast, structural cross-validation infers the comparability of models from the model structures and only applies the cross-validation tests hypothetically. The recovery submodel and a large part of the failure submodel are taken into account in the structural cross-validation part while other submodels are subject to the behavioural cross-validation tests.

Furthermore, some ordinary validation tests (Senge & Forrester, 1980) are applied to the aggregated model in section 4.3. Again, ordinary validation is divided in a structural part and a behavioural part. Regarding structural ordinary validation, the structure-verification test, the parameter-verification test, and the dimensional consistency test are applied. Regarding the structural-verification test, only structures are validated that are present in the aggregated model but not in the WSMD model. Other structures are covered with cross-validation tests. Note that the WSMD model is based on settled theory, therefore validation of common structures against the system of interest is not required. Regarding the parameter-verification test, only attention is paid to the numerical verification of uncertainty bandwidths of parameters.

Finally, behavioural validation tests concern the behaviour-reproduction test, the extreme-conditions test, and the boundary-adequacy test. The main difference between behavioural validation of the aggregated model versus the WSMD model concerns the number of runs to be evaluated (Auping, 2018). As the intended use of the WSMD model is consolidative in nature, only a single run could be evaluated. In contrast, the aggregated model provides an ensemble of runs, thereby facilitating a more thorough validation process.

4.1. Behavioural cross-validation

The behavioural cross-validation process is designed as follows. In the WSMD model, a single vessel and mission schedule configuration is chosen. As missions are aggregated to operations, and installations to installation groups, the cases with the highest heterogeneity in missions and installations in the WSMD model are naturally the most interesting. Therefore, the configuration is chosen in such a way that the usage rates of the installations and the mission effects have a high variance. Given the configuration, three scenarios are chosen; one with a low degradation rate, a normal degradation rate, and a high degradation rate. As running different configurations and different scenarios is quite time intensive with the WSMD model (due to the specification of many inputs in the database and a high running time), only a single configuration with three scenarios could be evaluated. Based on the WSMD model inputs, the aggregated model is initialized with use of the aggregation functions. Also, at the end of each run, the results of the WSMD model are aggregated in order to compare them with the results of the aggregated model.

The condition and the failure rate are subject to different tests. The condition outcomes will be cross-validated with the use of the behaviour patterns described by Yücel & Barlas (2011), while the failure rates will be cross-validated with the Root Mean Squared Error (RMSE) test. As the condition outcome will determine the shape of the failure rate, it is important to compare the condition course of the WSMD model with the condition course in the aggregated model. In that case, note that the condition chain in the aggregated model needs some sort of conversion function to map all the stages to a comprehensive condition stock. However, even with this mapping, computing the condition difference by means of the RMSE test is misleading, as the mapping of condition values to failures is non-linear (recap section 3.2.2). Therefore, the RMSE test will only be applied to the failure rate. Together, the tests will eventually indicate the degree of comparability between the condition, age, operations, and a part of the failures submodel. The simulated time of both models equals to 10 years (after this point, mission profiles are not specified by the WSMD database).

A final important note is that only a single behavioural pattern is present in the condition and the failure rate of the WSMD model. In the WSMD model, the condition exponentially decreases due to a reinforcing feedback-loop between the condition and the degradation of installations. However, as the recovery submodel is absent, no balancing feedback exists that results in a condition increase. Furthermore, due to the absence of renewals (this also belongs to the recovery submodel), failures in the infant phase cannot be simulated too. Due to the absence of infant failures and feedback in the condition, the failure rate only exhibits an exponential growth (with in the beginning a constant part, which represents the ‘normal’ phase in the failure generation).

4.1.1. Condition

The condition of the mechanical component class of the installation group immune to obsolescence (i.e. the NOR group) is shown in figure 4.1. Component classes, installations groups, and different degradation did not result in differences in behavioural patterns, so any combination could have been visualized. In general, the behaviour pattern of the aggregated model can be considered similar with the behaviour pattern of the WSMD model until condition 0.5. Until condition 0.5, the behaviour patterns of both models can be classified as a positive exponential decline. The exact slopes of the declines are dependent on the type of operation or mission, but even then it can be said that these degradation rates are proportional. For example, during 2021 and 2022, the slopes of the conditions flatten due to an overhaul. The positive exponential decline of the condition course in both models in no surprise due to the existence of a reinforcing feedback-loop from the condition to the degradation rate.

However, in contrast to the WSMD model, the aggregated model changes from a positive exponential decline to a negative exponential decline at condition 0.5. From this point, the difference between the use of a comprehensive stock for each installation and the use of a chain on a vessel level becomes clear. In the chain, component classes flow from stage to stage according to an average degradation time. This implies that component classes are distributed over multiple stages. In this light, the difference between the behaviour patterns of both models can be interpreted. Regarding the aggregated model, the majority of the component class at conditions below 0.5 are in reality in a lower stage, which results in a change of the behavioural pattern. However, the behavioural pattern of the component classes on an installation level remains the same, as the entire component class of an installation has the same condition.

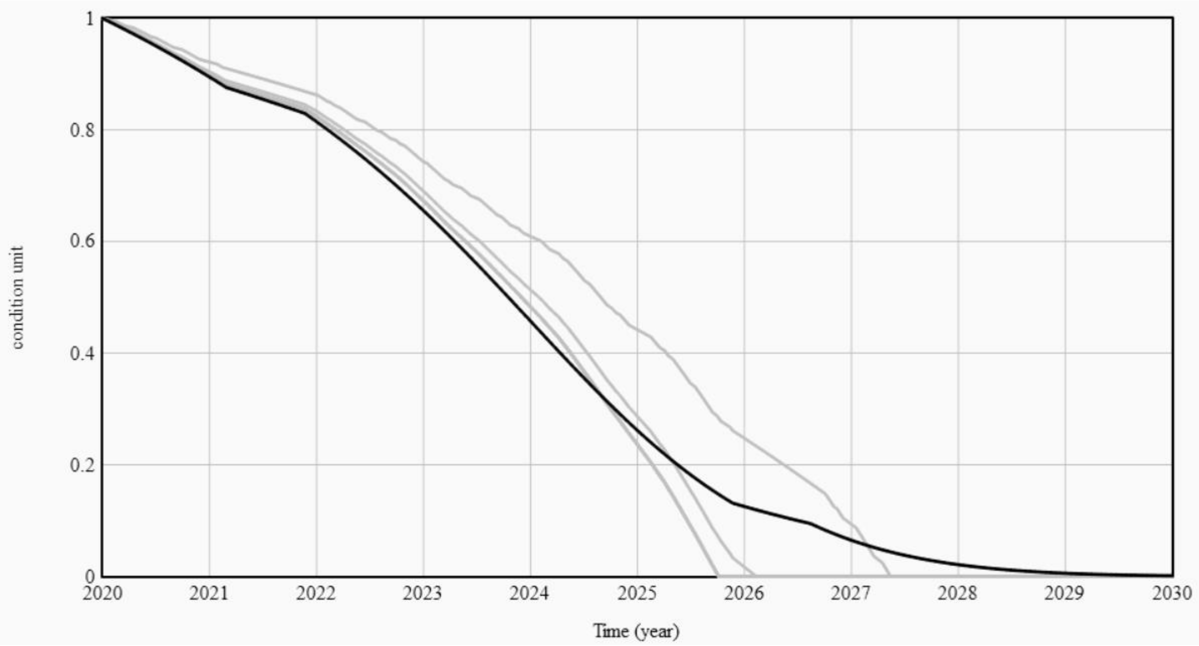

Figure 4.1. Condition outcome of the aggregated model (black line) and the WSMD model (grey lines) for a normal degradation scenario. The vertical axis shows the condition and the horizontal axis shows the time in years. Note that the three grey lines do not imply that only three NOR installations exist; multiple installations overlap. Also note that the aggregated model does not involve a comprehensive condition stock; the condition course in this figure is the result of a mapping of the condition stages to a comprehensive stock. This implies that even if the condition courses of both models were similar in this figure, they would result in different failure rates.

Now, what does this imply for the comparability of the models? As also communicated by the developers of the WSMD model, the use of a chain is in general a better approximation for modelling hierarchical objects. It should be kept in mind that in the WSMD model, aggregations are made of individual components which all may have a slightly different degradation rate. Therefore, it is more realistic in general to implement a model structure that accounts for the variation in higher resolutions than a model structure that assigns an equivalent condition to an entire object. Still, the change from a positive exponential decline to a negative exponential decline in the aggregated model is quite early. This may require a re-approximation of the condition chain, which will be further discussed in the next section.

4.1.2. Failure rate

The vessel failure rate for the scenario with the normal degradation rate is shown in figure 4.2. At first sight, the aggregated model may seem a quite accurate aggregation of the WSMD model. However, the area of interest only concerns failure rates between the lower dotted line and upper dotted line which represent 100 failures/year and 1000 failures/year respectively. Developing such a scoped definition of the consistency area is important in MRM, as not all outcomes are relevant for establishing the comparability between models (Bigelow & Davis, 2003). Regarding this case, the crew always can handle 100 failures/year while failures rates above 1000 failures/year could never be managed, even in the light of potential policies.

Considering the area of interest, the failure rates are quite different. For the average degradation scenario, the RMSE equals 218 failures per year for the error between the failure rates (i.e. the vertical error), and 1.07 year for the error between the timing of the failure rates (i.e. the horizontal error). As the vessel sailing availability is expressed as the realized sailing days divided by the planned sailing days, the horizontal error is deemed more important than the vertical one. Now, the degree to which errors can be tolerated, depends on the type of decision-making associated with the WSMD model and the aggregated model. Considering the facts that (a) the strategic timeframe in the aggregated model is in half years as sailing operations usually have a minimum duration of a half year, and (b) the tactical timeframe in the WSMD model is in months as missions usually have a minimum duration of a month, it is decided that horizontal errors above 0.4 year (approximately six months minus one month) are undesired. In this light, the models cannot be considered consistent with the current structure.

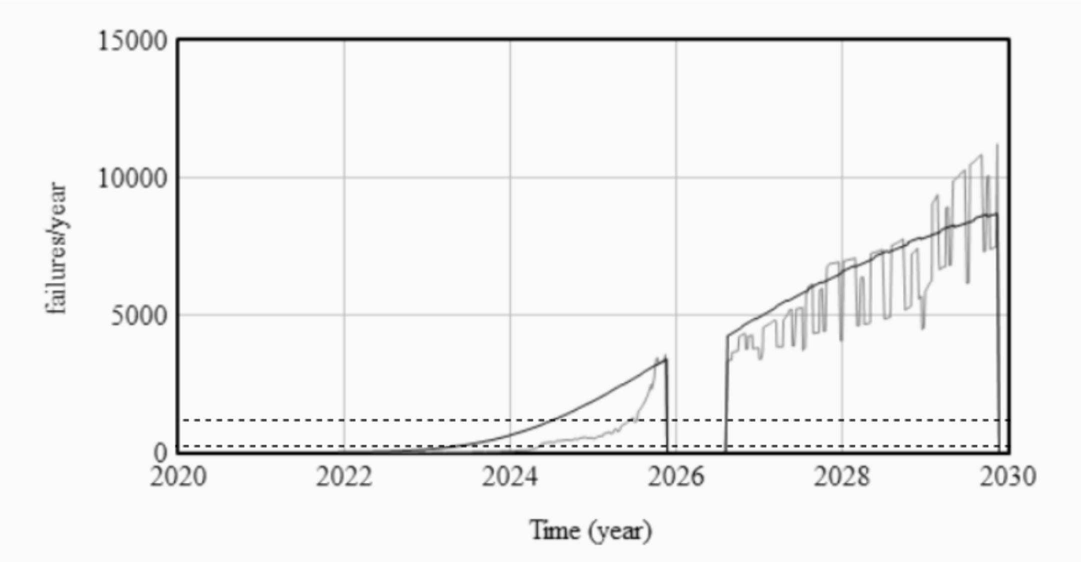

Figure 4.2. Failure rates of the aggregated model (black line) and the WSMD model (grey line) for a normal degradation scenario. The area between the dotted lines represents the failure rates of interest. The vertical axis shows the failures per year and the horizontal axis shows the time in years.

Given the errors in the failure rates for each scenario, it is concluded that in the aggregated model component classes are too widely distributed over the stages. This could be solved by increasing the number of stages. However, although this would result in a more accurate approximation, the computational complexity would also increase (recap the discussion in section 3.2.2). Another possibility would be to manually calibrate the models with a graphical function, which delays or speeds up the degradation rate of each stage. In order to prevent a higher running time, it is decided to implement the latter option. With the use of a calibration vector, the errors between the WSMD model and the aggregated model become much lower. For a comparison, please note table 4.1.

Table 4.1. Comparison of vertical errors and horizontal errors of the uncalibrated aggregated model and the calibrated aggregated model across different degradation scenarios.

	Vertical error (no calibration)	**Vertical error (calibrated)**	**Horizontal error (no calibration)**	**Horizontal error (calibrated)**
Low degradation	223 failures/year	32 failures/year	1.38 year	0.25 year
Normal degradation	218 failures/year	31 failures/year	1.07 year	0.21 year
High degradation	197 failures/year	22 failures/year	0.82 year	0.16 year

The question then remains whether a single calibration vector holds for different mission configurations. In order to prevent overfitting, another mission configuration of the WSMD model is chosen. In contrast to the previous configuration, installations now have similar usage rates, and missions come with the similar mission effects. Again, this configuration is run with a low, normal, and high degradation scenario. Table 4.2 compares the errors of the heterogenous configuration and the homogenous configuration (both with the use of a calibration vector). As can be noticed, the errors of the homogenous configuration are still below the horizontal error threshold of 0.4 year. The minor difference between the errors of the different configurations can be explained by the facts that (a) the aggregated model is able to capture a large share of the heterogeneity in installations by expressing them in component classes on a vessel level (recap section 3.2.2.), and (b) the heterogeneity in the mission profile only becomes visible at higher failure rates (recap figure 4.2) but these outcomes do not concern the area of interest for cross-validation. In this light, it can be concluded that for the cross-validated submodels in question, the calibration vector holds for different configurations.

Table 4.2. Comparison of vertical errors and horizontal errors of the calibrated aggregated model with a heterogenous (het) configuration and the calibrated aggregated model with a homogenous (hom) configuration across different degradation scenarios.

	Low (het)	**Low (hom)**	**Normal (het)**	**Normal (hom)**	**High (het)**	**High (hom)**
Vertical error	32 failures/year	29 failures/year	31 failures/year	28 failures/year	22 failures/year	28 failures/year
Horizontal error	0.25 year	0.38 year	0.21 year	0.30 year	0.16 year	0.22 year

4.2 Structural cross-validation

The absence of the recovery submodel in the cross-validation process may have implications for the comparability of the aggregated model and the WSMD model. Most importantly, the cross-validation tests could not sufficiently check whether the use of an equivalent value on an operations level could represent a variety of heterogenous missions in the WSMD model. Although the cross-validation process demonstrated that the heterogeneity in missions did not raise problematic consistency errors, this cannot be generalized to the uncross-validated submodels as well. For instance, the ability for maintenance heavily depends on the type of mission; therefore an aggregation function that maps this heterogeneity to a single maintenance ability for an operation may lead to intolerable errors when applying the RMSE test.

In case that initial aggregation functions result in intolerable errors (even with calibration vectors) after applying the cross-validation test, a possibility to capture heterogeneity on a high level is to create additional groups. This was done for most heterogeneities regarding installations; remember for instance from section 3.2.2 that the usage rates of installations were not aggregated to a single usage rate of an installation group but to a usage rate for each component class of an installation group. After this aggregation, the numerical error between the WSMD model and the aggregated model is tolerable. However, in some cases the creation of additional groups can be a bad idea. For instance, as also discussed in section 3.2.2, the abstraction level of the low-resolution model may become too high considering the intended model users. In other cases, the creation of additional groups is not possible as in that case one would end up at the resolution that required aggregation in the first place. For this study, this was the case for the obsolescence attribute of installations.

Still, the inability to capture heterogeneity in higher-resolution models is not necessarily problematic. In order to still achieve consistency, one could for instance decide to 'soften' the numerical error test by allowing a higher error. Another option would be to exclude the numerical error test and treat the entire range of high-resolution cases in the aggregated model as an uncertainty bandwidth. By treating aggregated factors as exploratory parameters, the aggregated model can highlight the areas of interest for the WSMD model with the use of exploratory analysis. For instance, if the operational harshness of missions in the WSMD model would not be numerically consistent with the operational harshness of operations in the aggregated model, the aggregated model can still highlight 'problematic' missions by exploring the entire external impact range. An illustration of this is provided in section 5.2.

For this study, certain obsolescence attributes of installations are not included in the WSMD model. As stated in section 3.3.3., the creation of additional 'obsolete' groups was heavily inefficient. As it was decided to maintain the numerical cross-validation test, the only option left was to accept that the aggregated model could not capture all information deemed important the WSMD model. In this light, only a single obsolescence type is used in the aggregated model. Due to this choice, the number of insolvable failures due to other types of obsolescence cannot be estimated with the aggregated model. As a consequence, it is not possible to simulate mission profiles with an interval between two consequent overhauls longer than the minimum obsolescence free period of the installation set. This results in a maximum overhaul interval of around 6 years. Fortunately, the maximum operational duration of the vessel in question is not longer than 6 years, so this is not a huge drawback of the aggregated model. However, it is important to note that for an overhaul interval longer than 6 years, it is not likely that the models are comparable with respect to their sailing availability.

4.3. Ordinary validation

In general, most validation tests applied to the aggregated model were passed. The validation of new structures in the aggregated model (i.e. the CBM structure and the maintenance priority structure) was done in collaboration with experts associated to the Dutch force support system, as the structures were very case specific. Furthermore, the associated bandwidths were developed on basis of expert opinions and scientific literature.

Regarding the dimensional consistency test, no unit errors were encountered. Finally, after the extreme conditions test, no conditions exist that are able to break the aggregated model. The structural validation tests and the extreme-conditions test are more elaborately described in appendix A. Regarding the other behavioural validation tests, some important observations are shared. The base ensemble of the aggregated model for some important outcomes is shown in figure 4.3. The experimental setup of the base ensemble is stated in chapter 5.

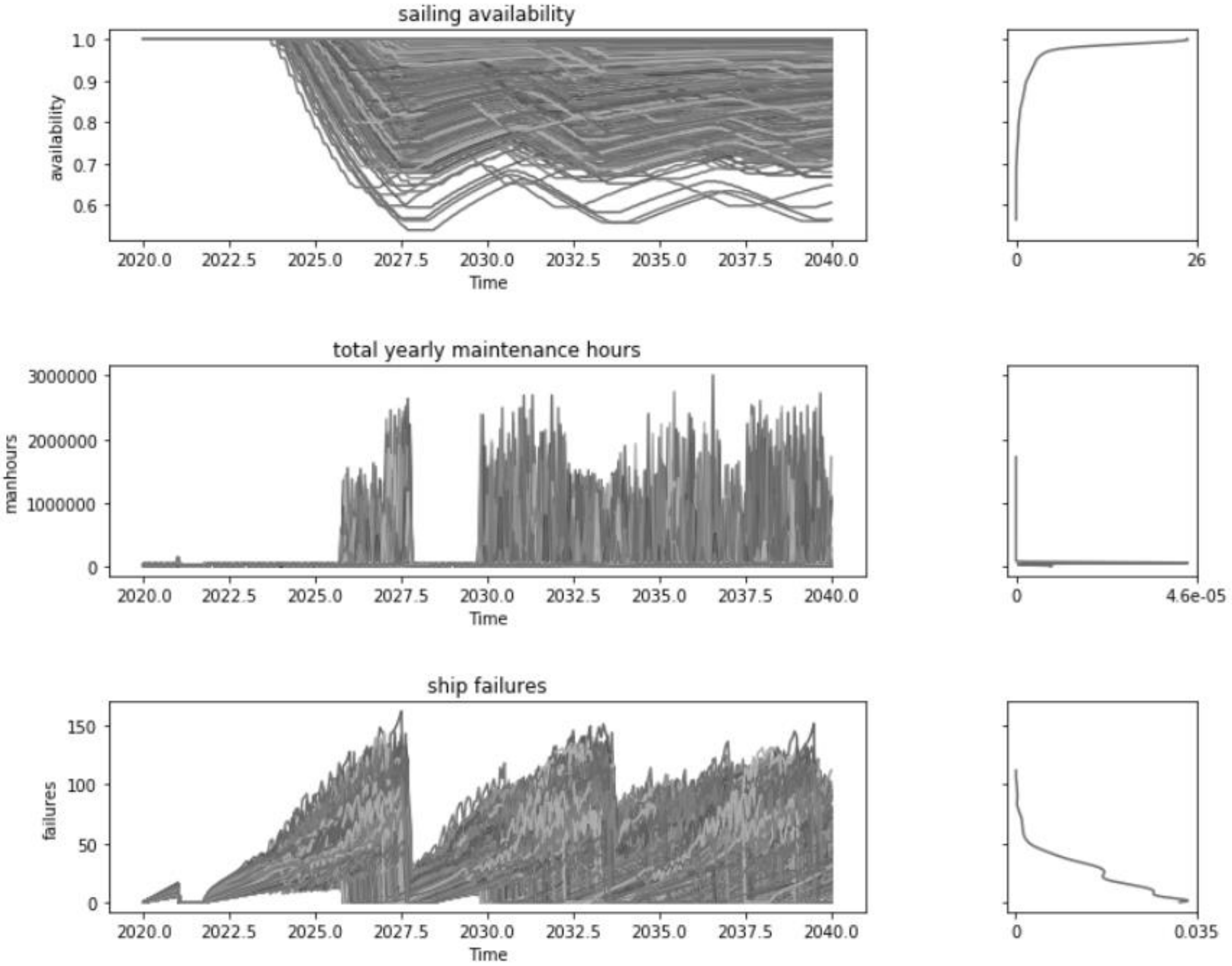

Figure 4.3. Base ensemble (5000 scenarios) of the aggregated model. The figures on the right represent the density of the ensembles at the final time step of the simulation. The time between consequent overhauls and the number of stops for intermediate maintenance are included as a parametric uncertainty. Also, the CBM structure is included as a structural uncertainty. During robustness analysis in chapter 5, these will be treated as levers.

First, regarding the behaviour-reproduction tests, it appears that scenarios exist where an increasing failure rate cannot be managed by the crew, thereby leading to a decreasing sailing availability and an increasing workload due to overhaul. The increasing workload during overhaul can be derived from the high peaks at the yearly maintenance hours outcome. As communicated by the WSMD model developers, such outcomes sometimes occur in reality but are often not taken into account by decision-makers and reliability engineers as they tend to base the design of the force support system on scenarios that are deemed 'likely'. Chapter 5 provides more information on worst-case scenarios and consequent system interventions that may potentially close a sailing availability gap.

Regarding the boundary adequacy test, the base ensemble of the aggregated model highlights some scenario spaces that may be decrease the validity of the WSMD model. First, in the WSMD model, it is assumed that the amount of work during overhaul can be completed in a fixed time. As the structure of the aggregated model is based on the WSMD model, the aggregated model has a short high peak at the overhaul intervals, implicitly assuming that suck peaks can always be managed within the available time (the short peaks are difficult to see in figure 4.3 as the time between overhauls is made variable, leading to many peaks in the ensemble).

As communicated by the WSMD model developers, in most cases the overhaul crew hires additional human resources to finish the required amount of work in time. Still, it would be better to make the duration of overhaul dynamic by including a balancing feedback-loop involving the condition of the vessel at arrival (which indicates the amount of work to be done) and the available human resources. When including this structure, cases may occur where a high amount of work during overhaul results in a delay of the overhaul operation which decreases the sailing availability. Now, although the aggregated model advocates a 'boundary move', the implementation of a HR system remains a suggestion of further research.

A second potentially problematic observation that resulted from the boundary adequacy test concerns the fact that the WSMD model assumes a constant productivity of the crew. However, the question is whether this holds in the light of scenarios that require a higher workload than initially expected. Actually, the phenomenon of 'diminishing returns' is known to be one of few economic theories that might be called a law (Shephard & Färe, 1976). For the WSMD model, a model structure where diminishing returns can become dominant concerns the corrective maintenance part; illustrated in figure 4.4.

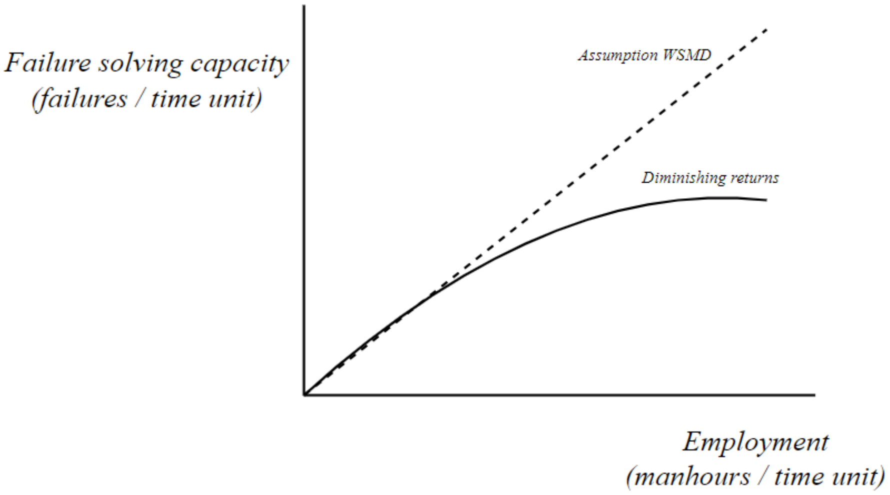

Figure 4.4. Illustration of relationship between the failure solving capacity and the employment on a vessel. This relation will eventually always hold, however the question is whether diminishing returns are already relevant at current employment levels.

Concrete, diminishing returns become important when installations have not sufficient capacity to simultaneously solve a certain cumulative amounts of failures. Currently, the WSMD model assumes a constant productivity (where the productivity is the derivative *dy/dx* in figure 4.4). However, for a variety of reasons, ranging from technical challenges (some failures need to be sequentially solved) to practical issues (i.e. an installation only has limited physical space for crew members to solve failures), the productivity of the crew can decline. Now, some scenarios result in cumulative failures that are likely to disprove the constant productivity assumption. To illustrate, many scenarios achieve cumulative failure amounts of above 100 failures. At best, given a dozen of installations, this would imply that installations allow a simultaneous solve of approximately 10 failures. In reality, this is likely unrealistic and the productivity of the crew will decline. However, similar to the fixed overhaul time, the aggregated model does not implement a balancing feedback-structure due to the limited available research time.

In response to the limitations of the aggregated model and the WSMD model that resulted from the boundary adequacy test, a high uncertainty bandwidth is implemented for the crew productivity in general. Still, the absence of a balancing feedback-loop from the failures to the productivity of the crew can be problematic in case that decision-makers aim evaluate policies to eliminate an availability gap. When no care is taken, decision-makers may 'push' the system in the wrong direction. To illustrate, an increase in the available working hours of a crew is useless when current employment levels do not lead to a higher failure solving capacity. Chapter 5 takes this into account when performing a robustness analysis.

4.4. Key takeaways of validation

- Model validation in MRM is a two-way process.

    1. As aggregations lead to information loss, the aggregated model requires cross-validation by the WSMD model in order to determine its fitness for purpose.
    2. Vice versa, as the type of decision-making associated to the WSMD model needs to deal with high uncertainty, the exploratory use of the aggregated model has implications for the validity of the WSMD model

- The aggregated model has different consistency standards than the WSMD model.

    1. The aggregated model does only require consistency with the WSMD model in a specific outcome area of the failure rate. Areas outside the area of interest are deemed irrelevant as they do not affect the sailing availability outcome.
    2. As the use of the aggregated model is more strategic in nature, the aggregated model allows a higher numeric error with real-world data than the WSMD model.

- In order to achieve a tolerable consistency error between the WSMD model and the aggregated model, the chain in the aggregated model required calibration. The calibration vector holds for different initializations of the WSMD model.

- Given the cross-validation tests, finding proper aggregation functions for the obsolescence structure in the WSMD model was found impossible (even when including calibration vectors or increasing the number of aggregated groups). As it was decided to maintain the cross-validation tests, a large part of the obsolescence structure is omitted in the aggregated model. This implies that models are not consistent when overhaul intervals are longer than 6 years. This is deemed acceptable, as the intended overhaul intervals for the vessel in question are not longer than 6 years as well.

- By means of exploratory analysis, the aggregated model found some potentially problematic assumptions made by the WSMD model. More concrete, the assumption of a fixed overhaul duration and a constant crew productivity are not likely to hold in the light of scenarios with an unexpectedly high cumulative workload.

5. Results

The case in this chapter concerns the question whether the current maintenance plans of the Dutch Royal Navy as depicted by the WSMD model are sufficient to facilitate a robust sailing availability of a specific vessel given varying durations between two consequent overhauls (hereinafter referred to as an overhaul interval). The vessel in question currently executes mission profiles with an overhaul interval no longer than four years. In the light of increasing expectations of international safety & security organisations, a strong aim exists to increase the overhaul interval to five years or even six years. However, the question is whether the current maintenance plans support an extension of the overhaul interval. Next to living up of international expectations, also financial motives exist for this analysis. A few months ago, the commanders of the Dutch armed forces advocated for a significant budget increase in order to keep protecting friendly territories and to maintain the international legal system and stability (Trouw, 2021). This analysis may facilitate a discussion about the urgency of this problem. Now, the problem is further conceptualized with the XLRM framework.

5.1. Problem conceptualization

The XLRM framework is a conceptualization tool for problems that deal with massive uncertainty (Lempert et al., 2003). XLRM involves the conceptualization of uncertainties (X), levers (L), the simulation model (R), and the outcomes (M). The uncertainties and levers are more elaborately described in appendix B.

X. Each submodel comes with a few uncertainties, resulting in 15 parametric uncertainties in the aggregated model. This is a huge difference compared to the number of uncertainties in the WSMD model, which were at least a tenfold. In the condition submodel, most uncertainties are located in the degradation part. In the degradation part, the uncertainties include the degradation due to time, degradation due to usage and degradation due to external factors (e.g. the area where the operation takes place). In the failure submodel, uncertainty exists around the initial MTBF of component classes and the consequent decrease in performance of a vessel. Uncertainties in the maintenance submodel primarily relate to the required manhours to do preventive maintenance, to execute a renewal, or solve a failure. In the operations submodel, the only uncertainty concerns the allowed weeks per year for intermediate maintenance, which is often determined by external parties. Finally, in the age submodel it is uncertain to what extent the age impacts the maximum achievable condition of the vessel.

L. The current maintenance plans as depicted by the aggregated model primarily depend on the time and the usage. These plans are aggregated from the WSMD model and will be evaluated with robustness metrics. Next to these maintenance plans, the aggregated model also takes into account condition-based maintenance (CBM), a maintenance type that is modelled in the aggregated model only (please recap section 3.3.4). After a vulnerability analysis, the robustness of the maintenance plans are compared to a variety of additional policies for an overhaul interval of 4 years, 5 years, and 6 years. These alternative policies are covered in section 5.3.

R. Needless to say, the simulation model in question denotes the aggregated model. Table 5.1 shows the main simulation and exploration settings.

Table 5.1. Simulation settings

Time step	Simulation time	Integration method	Number of runs	Number of repetitions	Sampling method	Sampling distribution
1/128 year	20 years (2020 – 2040)	Euler	The base ensemble* contains 5000 scenarios. Policies are separately evaluated with 500 scenarios per policy.	0, as the model is deterministic	Latin Hypercube	Uniform

* The base ensemble is used in the validation chapter (section 4.3) and will also be used for vulnerability analysis (section 5.2)

M. Regarding the outcomes, the sailing availability and the working hours are of interest (please recap section 3.3.6 for the specification). Regarding the availability, asset owners usually have a ‘Service Level Agreement’ (SLA) with a maintenance organisation that states the minimum guaranteed performance of an asset (Accenture, 2017). Regarding this case, it is assumed that the Dutch Royal Navy agrees to an availability of 95% for an overhaul interval of 4 years, meaning that the current budget only allows a discrepancy of 5% between the planned sailing time and the realized sailing time. However, as SLAs are usually based on a few mission profiles only, it is unclear whether the current maintenance plans also achieve a 95% availability across a wider range of scenarios. Moreover, it is unknown what happens when the overhaul interval increases to 5 years or 6 years, or whether higher SLAs are possible. Regarding the working hours, the working hours in the aggregated model merely serve as a constraint for keeping the vessel operational; therefore minor deviations will be deemed less interesting by the Dutch Royal Navy as the working hours in an overhaul interval cannot exceed the capacity.

In order to quantify the performance of the levers on the outcomes, two robustness metrics are chosen. First, as decision-makers are quite risk averse towards the sailing availability, Starr's domain criterion (Schneller & Sphicas, 1983) is chosen as a robustness metric. In short, Starr's domain criterion (SDR) can be classified as a satisficing robustness metric (McPhail et al., 2018) which (a) sets a harsh threshold that scenarios must satisfy, (b) transforms availability outcomes to binary values according to whether scenarios satisfy the criterion, and (c) computes the mean of the transformed availability outcome of the scenario set. Starr's domain criterion fits well a risk averse attitude, as scenarios that have an availability slightly lower than the SLA are punished, while they are generally favoured by the more statistical metrics (e.g. mean-variance metric, percentile-based metrics etc.) (McPhail et al., 2018).

Second, regarding the working hours outcome, the mean-variance metric (Hamarat et al., 2014) is chosen. The mean-variance metric aims to capture the overall system performance by expressing the robustness of a policy as a function of the mean and the standard deviation of the outcome of interest. This metric is less risk averse than Starr's domain criterion, as slight differences in the mean or variation of the working hours are generally not punished. All in all, the metrics with regard to the outcomes can be formalized with equation 4 and 5 for the mean-variance metric and Starr's domain criterion respectively. For both outcomes, the robustness metrics only use the outcome values at the final time step of the simulation.

$$MV\_wh_{j,t} = \mu\left(\overrightarrow{wh}_{j,t}\right) * \sigma\left(\overrightarrow{wh}_{j,t}\right) \tag{5.1}$$

$$\begin{gathered} SDR\_a_j = \frac{1}{N}\sum_{s=1}^{S} a'_{j,s} \\ where\ a'_{j,s} = \begin{cases} 1\ if\ a_{j,s} \geq SLA \\ 0\ if\ a_{j,s} < SLA \end{cases} \end{gathered} \tag{5.2}$$

Regarding the notation, $\overrightarrow{wh}_{j,t}$ represents a vector containing the working hours of type *t* (either maintenance or renewal) of all scenarios *S* for policy *j*. Furthermore, $a'_{j,s}$ represents the reward of $a_{j,s}$ which in turn represents the availability outcome of scenario *s* given policy *j* and the SLA. *N* denotes the number of experiments.

5.2. Vulnerability assessment maintenance plans

The vulnerability analysis demonstrated that a linear increase in the overhaul interval cannot be managed with a linear increase in the working hours. Figure 5.1 visualizes the impact of the main parameters that make up the scenarios with a lower availability than 95%. In other words, when a scenario does violate the current SLA, it is considered a worst-case. Regarding the figure, the colour scale represents the number of worst cases divided by the number of total cases. The '0-1-2' range represents the first, second, and third 33.33% of the parameter bandwidth in question. For example, the 0, 1, and 2 of the overhaul interval represent the range 4 – 4.67 years, 4.67-5.33 years, and 5.33-6 years respectively.

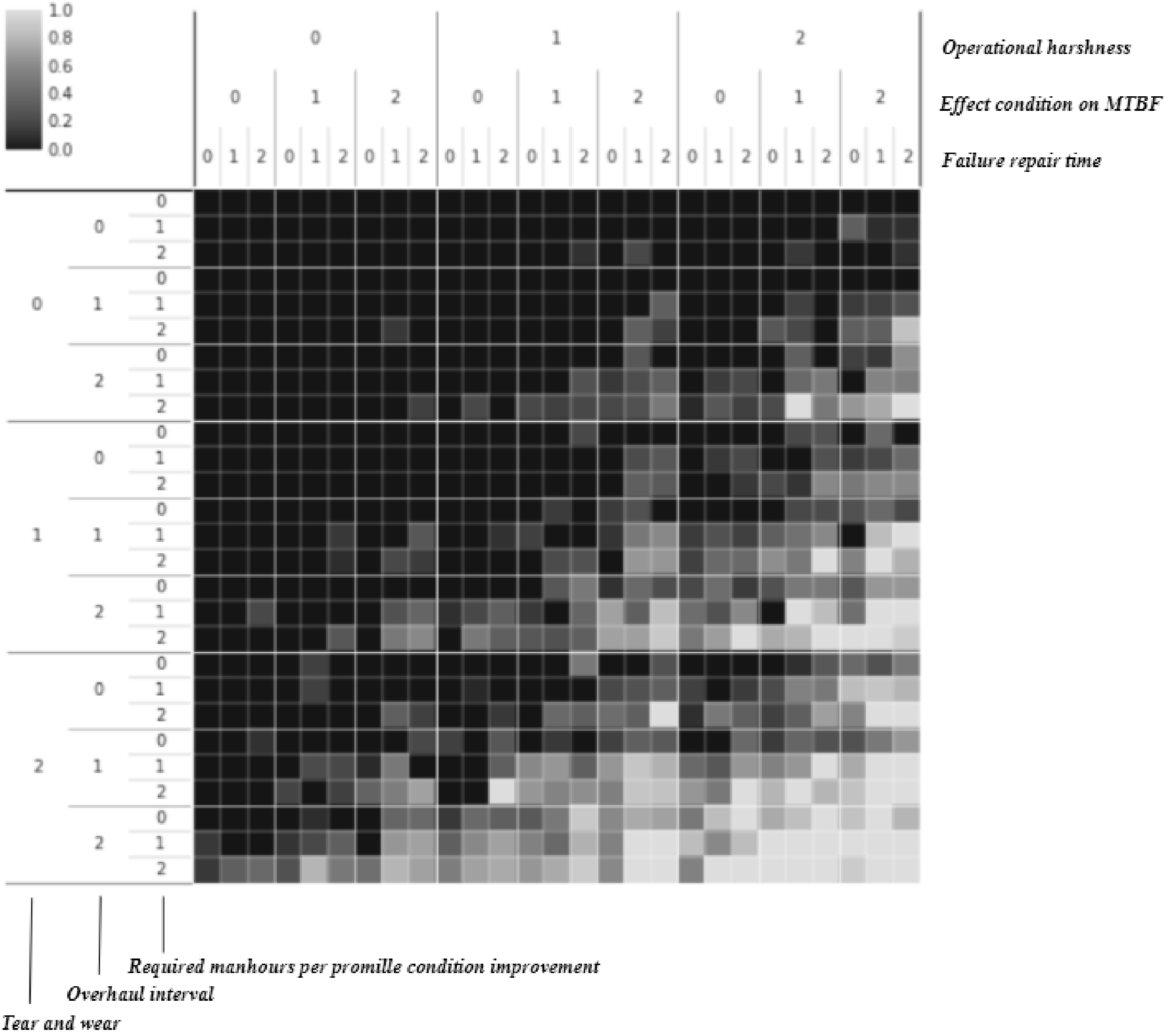

Figure 5.1. Visualization of worst-cases regarding the availability outcome with dimensional stacking (Suzuki et al., 2015). The colour scale represents the number of worst cases divided by the number of total cases (5000). The '0-1-2' range represents the first, second, and third 33.33% of the parameter bandwidth in question.

Based on scenario discovery algorithms (Friedman & Fisher, 1999; Suzuki et al., 2015) three problematic scenarios are identified.

1. Scenario *'high degradation'*. Scenarios that are characterized by a high degradation (due to tear & wear and the operational impact) dramatically increase the number of worst-cases. More in-depth, PRIM finds ranges of 0.11 to 0.14 and 0.36 to 1 for the tear & wear and the operational impact respectively to be problematic. The interpretation of these ranges is confidential.
2. Scenario *'long overhaul interval'*. PRIM finds the increase in the overhaul interval to 5 years or more significantly contributing to the number of the worst-cases. Note that the current overhaul interval of the vessel in question denotes 4 years.
3. Scenario *'low productivity'*. Scenarios become 'quicker' a worst-case when the productivity of the crew turns out to be lower than expected. In particular, a failure repair time between 38 hours and 75 hours is found to be problematic by PRIM.

Based on dimensional stacking, it is concluded that the high degradation scenario has the most impact on the number of worst-cases. The low productivity scenario has the weakest impact on the number of worst-cases. Obviously, the occurrence of all scenarios together will be most problematic.

All scenarios imply that the relation between the overhaul interval and the outcomes of interest is not linear and that additional effort is needed to increase the overhaul interval. This stands to reason as components within installations affect each other's condition, thereby creating exponential degradation behaviour. Therefore, the condition of a vessel does not linearly decrease when the overhaul interval linearly increases. Due to exponential behaviour, a 'tipping point' occurs when the maintenance crew cannot provide the required amount of maintenance to keep the condition optimal (due to a high degradation and/or a low productivity). This does in principle not need to be problematic, as a vessel has some in-built redundancy which 'delays' the interval between the tipping point and the moment when the vessel becomes unavailable during a sailing operation. However, in case when the overhaul interval increases, this buffer often appears to be insufficient.

The degree of influence of decision-makers on the scenarios differs. In general, the tear and wear factor is outside the influence sphere of decision-makers, as this reflects the natural degradation of objects over time. Furthermore, the operational harshness represents the impact of missions in the WSMD model, which is for this study assumed to be not under control of decision-makers. This is also the case for the overhaul interval. Finally, decision-makers have the ability to increase the productivity of the crew by for instance providing more training services or by improving the design of objects. Now, the next section will evaluate the robustness of the maintenance plans and also elaborate on potentially promising alternative policies.

5.3. Robustness maintenance plans

For both the initial plans as condition-based maintenance (CBM), achieving a 95% sailing availability for operational durations of 4 years, 5 years or 6 years is not possible. Figure 5.2 visualizes Starr's domain criterion for multiple service level agreements. With both maintenance plans, the current SLA can be achieved for approximately 95%, 85%, and 75% of the scenarios for an overhaul interval of 4 years, 5 years, and 6 years respectively. In general, achieving higher SLAs exponentially increases the required effort as the sailing availability exponentially decreases.

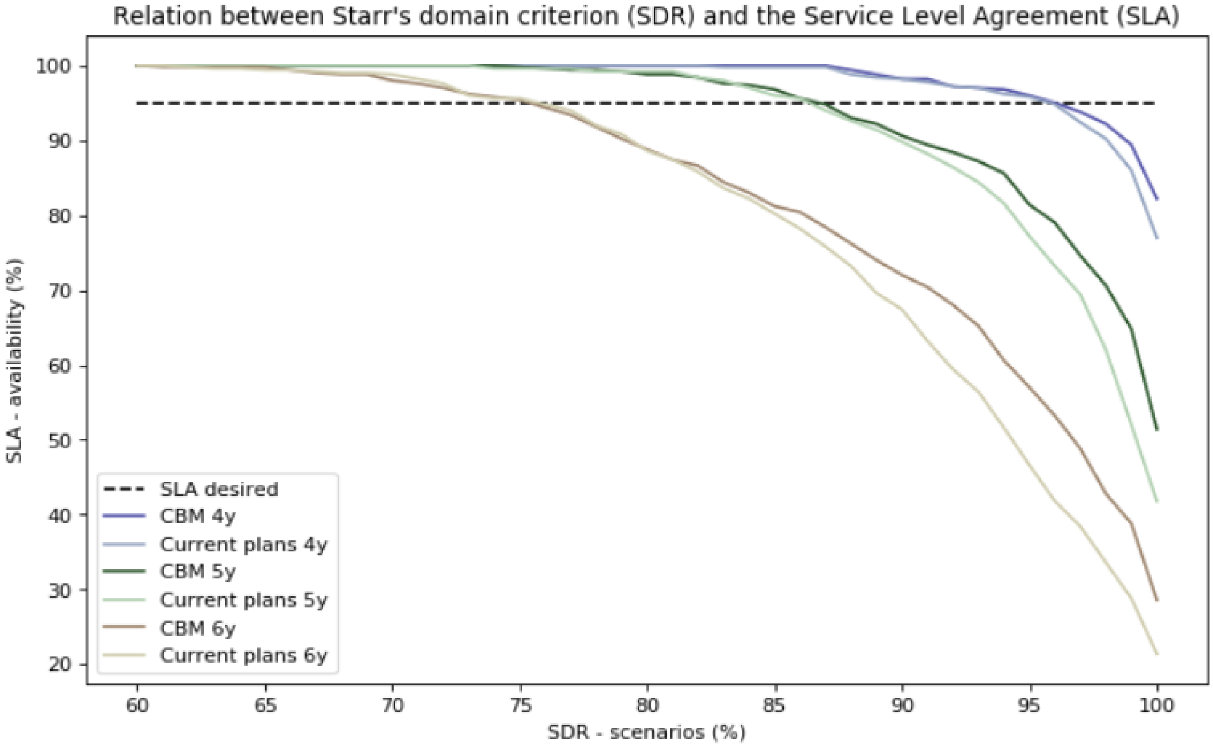

Figure 5.2. Starr's domain criterion plotted against multiple service level of agreements for the maintenance plans (500 scenarios per planning type). The dotted line represents the current SLA of 95%.

Regarding the robustness of the maintenance plans, note that the initial maintenance plans have a worse performance than CBM when shifting the SLA to lower availabilities. The reason for this is that CBM provides a more adequate reaction on degradation behaviour; the lower the condition of an object the higher the assigned preventive maintenance hours. The better performance of CBM than static maintenance plans is not an uncommon finding (Sharma et al., 2017; Esmaeili et al., 2019). Still, CBM does only slightly improve the sailing availability and the working hours. This can be explained by the fact that the current maintenance plans are already quite near the maintenance capacity which cannot be exceeded.

Another important observation is that an increase in the overhaul interval exponentially increases the required effort. This can be derived from figure 5.3 which visualizes the mean-variance scores for the working hours. A higher score denotes a lower robustness. As can be seen, the scores for the renewal hours exponentially increase when the overhaul interval increases. This stands to reason, as a longer overhaul interval leads to a lower condition of objects, thereby increasing the number of renewals. The scores for the maintenance hours remain more or less constant, as the maintenance capacity cannot be exceeded. Now, although CBM does not ‘outperform’ the current plans, it can be stated that CBM (despite its higher fixed monitoring & inspection hours) has a better performance than the current plans with respect to the working hours. Combined with a better robustness performance for the availability outcome, it can be concluded that CBM proves to be a more robust maintenance plan than the current ones.

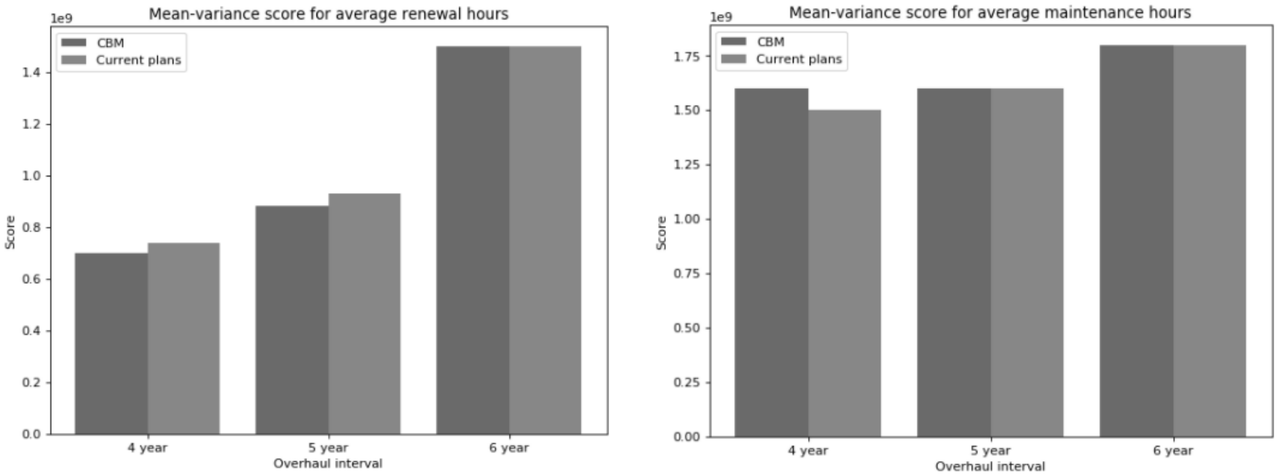

Figure 5.3. Mean-variance score for the working hours (based on 500 scenarios per policy). The higher the score, the lower the policy robustness

However, when individually implemented, CBM is insufficient to achieve an SLA of 95%. For an overhaul interval of 4 years, the robustness of the sailing availability might be deemed sufficient. However, for 5 years or even 6 six years between two consequent overhauls, decision-makers may not be content with the robustness curve for the sailing availability. In this light, six additional policies (together with CBM) are implemented and evaluated. Two policies concern an increase in the available maintenance hours; one policy assigns a part of the intermediate maintenance crew to the sailing crew and one policy increases the available maintenance capacity for all crews. Next, a policy is implemented that increases the number of stops at the harbour for intermediate maintenance (without adjusting the total intermediate maintenance effort per year). Furthermore, a policy concerns an increase in the condition threshold for renewals and another one increases the redundancy of objects. Finally, a policy is implemented that increases the productivity of the crew. The robustness of these alternatives is described in the next section.

5.4. Robustness alternatives

Before discussing the performance of the alternatives, an important note should be made about the validity of the results. In short, it is too early to state a comprehensive ‘policy ranking’ based on the differences between the performance of alternatives. As such ranking would require complete cross-validation in the future, the current aim is to indicate the ‘sensitivity’ of the outcomes to policies rather than stating which alternative performs ‘best’. Stating which policies are most optimal also requires a more sophisticated procedure of policy specification and evaluation. Furthermore, it is misleading to state the performance of the policies that increase the available working hours of a crew, as a higher assignment of manhours to maintenance activities will likely lead to diminishing returns (recap section 4.3). Based on this, it is also difficult to compare these policies with the productivity increase policy. Still, the current analysis can provide helpful insights in which policies are worth considering when designing a robust force support system. Now, figure 5.4 visualizes the robustness scores for the sailing availability outcome, and table 5.2 contains the robustness scores for the working hours outcome.

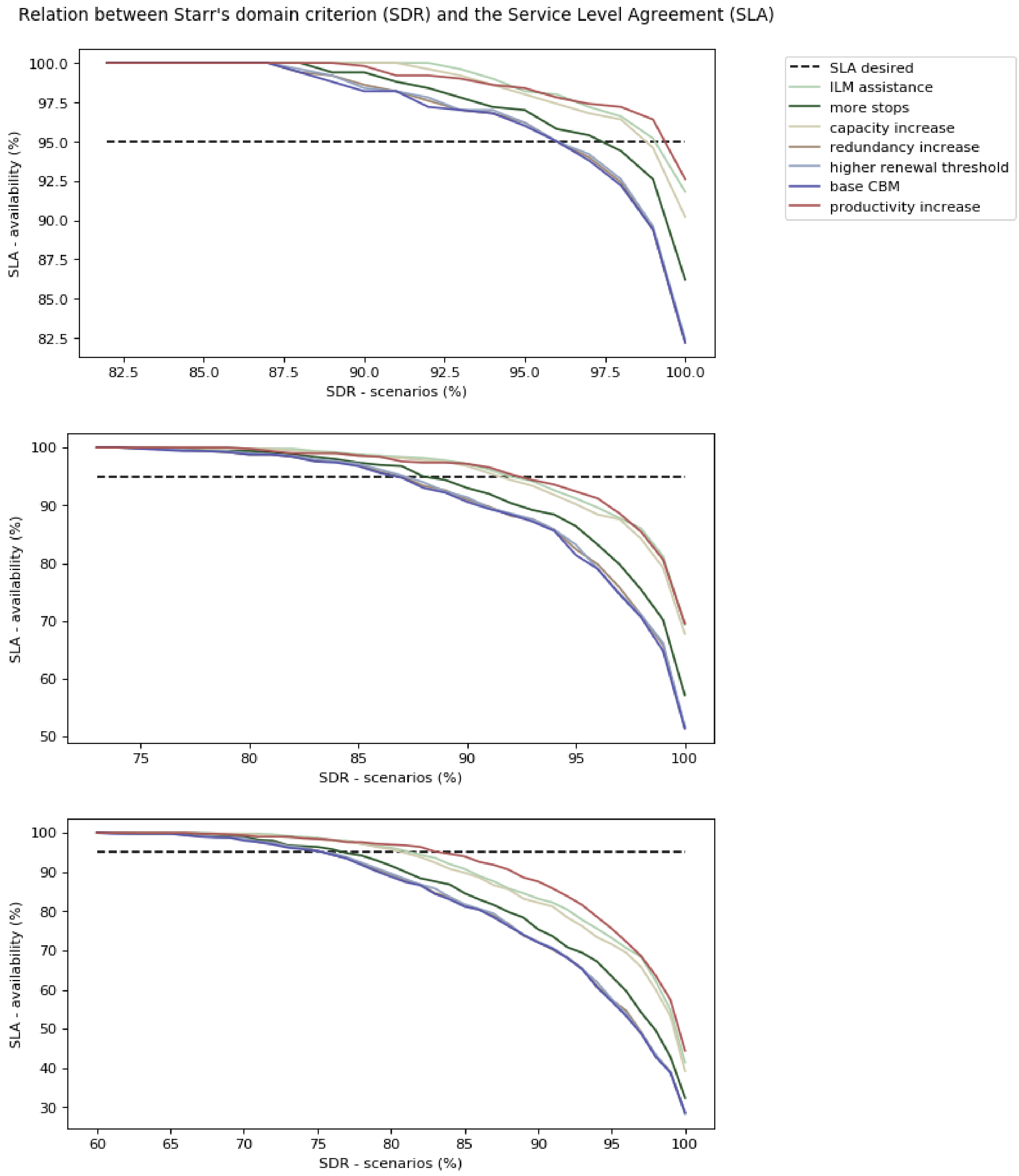

Figure 5.4. Starr's domain criterion plotted against multiple service level of agreements for multiple overhaul intervals (based on 500 scenarios per policy). The dotted line represents the current SLA of 95%. The upper, middle, and lower graph represent an overhaul interval of 4 years, 5 years, and 6 years respectively.

Table 5.2. Mean-variance robustness scores for the average maintenance costs per operational year (TMC_a) and the average renewal costs per overhaul (TRC_a). The higher the score, the lower the robustness. Scores are in billions.

Outcomes **Policies**	*TMC_a 4 years*	*TRC_a 4 years*	*TMC_a 5 years*	*TRC_a 5 years*	*TMC_a 6 years*	*TRC_a 6 years*
Base CBM (alternative zero)	1.6	0.7	1.6	0.9	1.8	1.5
ILM assistance	1.6	0.6	1.6	0.8	1.8	1.3
More IM stops	1.6	0.7	1.6	0.9	1.8	1.5
Capacity increase	2.3	0.6	2.4	0.8	2.5	1.3
Redundancy increase	1.6	0.9	1.6	1.1	1.8	1.8
Higher renewal threshold	1.5	0.7	1.6	0.9	1.8	1.6
*Productivity increase**	1.6	0.7	1.6	0.9	1.8	1.5

**Costs for the productivity increase were not included in the model design, as an increase in the productivity can be facilitated in a variety of ways. Therefore, the costs are equal to the CBM costs.*

To start with, the policy to increase the condition threshold for renewals and the policy that increases the redundancy of objects both perform poorly. Both policies result in a minimal improvement of the sailing availability (compared with the 'zero' alternative, which is the implementation of CBM only) while they come with relatively bad scores for the renewal working hours. These results imply that some sort of 'buffer', either by early-stage renewals or additional redundancy, does prove to be ineffective when implemented outside the overhaul interval. In other words, when policies do not deal with exponential degradation behaviour during sailing operations, they are not likely to result in a good performance.

A slightly better performing policy is the increase in the amount of stops at a harbour for intermediate maintenance. With this policy, the exponential degradation loop is more often interrupted, thereby facilitating a delay in the condition deterioration and an increase in the sailing availability. An additional advantage of this policy is that it does not come with additional maintenance and renewal hours. However, it should be noted that it is not always possible to increase the number of stops, especially in the light of demanding missions that barely allow a 'break' of the vessel in question.

At first sight, policies that increase the working hours rate of the crew (i.e. the policies 'ILM assistance' and 'capacity increase') have a way better availability score than 'Base CBM'. However, as described in the start of this section, it is not possible to state the effectiveness of these policies due to absence of the 'diminishing returns' feedback-loop. Still, both policies can be compared with each other. In fact, both policies seem to have a similar performance on the sailing availability, but have different scores on the working hours. Concrete, when one aims to choose between these two policies, assigning a part of the ILM crew to the sailing crew will lead to more robust outcomes on the working hours outcome than increasing the size of both crews.

Finally, it can be stated that an increase in the productivity appears to be an effective policy in robustly increasing the sailing availability of the vessel in question. Just like the policies discussed in the previous paragraph, this policy creates more available working hours. However, unlike the policies in the previous paragraph, this policy does not involve possible diminishing returns as the initial size of the crews do not change. The high effectiveness of this policy did not come as a surprise, as section 5.2 already pointed out that a 'low productivity' scenario had significant impact on the number of worst cases. Still, it should be noted that the costs of this policy is not included, as a higher productivity can be achieved in a variety of ways. This evaluation of the costs should first be evaluated before considering the implementation of this policy.

5.5. Key takeaways of results

- By stress testing the maintenance plans on a wide range of scenarios, the aggregated model demonstrated that a linear increase in the overhaul interval (i.e. the time between two consequent overhauls) cannot be managed with a linear increase in the working hours.

- Scenarios characterized by a high degradation, a long overhaul interval, and/or a low crew productivity significantly decrease the performance of maintenance plans with respect to the sailing availability and the working hours.

- In general, robustly achieving higher SLAs or longer overhaul intervals requires an exponential increase in the working hours. This can be explained by the exponential degradation behaviour of objects due their lower-level interactions.

- Of all maintenance plans, condition-based maintenance (CBM) achieves the most robust outcomes. Still, when increasing the overhaul interval to 5 years or more, the implementation of CBM only likely leads to weak outcomes. In this light, it is promising to further investigate the real-world effects of an increase in the productivity of the crew as this potentially leads to a higher robustness of the force support system.

6. Discussion

This chapter discusses some topics that require more attention in the MRM field. First, the consistency framework used in this study is evaluated. Second, the role of cross-validation tests with regard to the usefulness of models at different resolutions is discussed. Finally, the role of MRM in industrial applications is discussed.

6.1. Consistency framework evaluation

A point of discussion regarding the consistency framework used in this study concerns the use of aggregation functions. During the validation process, it was noticed that the aggregations functions from the WSMD model initial state to the aggregated model initial state impeded the quality of cross-validation. The main reason for this is that the WSMD model only allowed a limited number of runs, thereby hindering a proper search of the input space. In line with Bigelow & Davis (2003), it would be better to use the consistency framework in figure 6.1 where the initial state of high-resolution model is initialized with a disaggregation function.

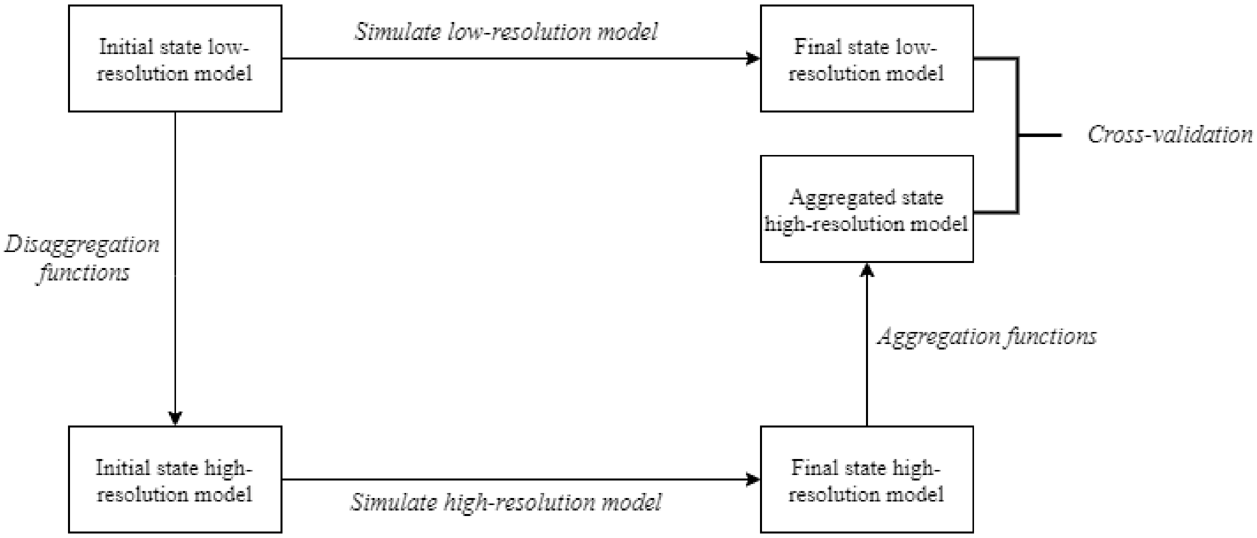
Initial state low-resolution model
Simulate low-resolution model
Final state low-resolution model
Cross-validation
Aggregated state high-resolution model
Disaggregation functions
Aggregation functions
Initial state high-resolution model
Simulate high-resolution model
Final state high-resolution model

Figure 6.1. Improved consistency framework. Adapted from Bigelow & Davis (2003)

The disaggregation of the low-resolution inputs to high-resolution inputs facilitates a more thorough cross-validation process. As the low-resolution model is suitable for exploratory analysis, the low-resolution model can be used to explore the uncertainty space of the system of interest. Concrete, the low-resolution model may identify a number of 'behavioural clusters' that make up the outcome space of interest (Steinmann et al., 2020). Consequently, an illustrative scenario run can be selected from each cluster and disaggregated to high-resolution inputs by use of a disaggregation function. This search for 'exemplars' that capture the outcomes of interest was not possible with the WSMD model.

In case of disaggregation, it should be noted that a number of different high-resolution parametrizations may be needed for each low-resolution state. As discussed multiple times in this study, the heterogeneity in a high-resolution initial state matters for the results of the models. For instance, a similar usage rate for installations in the WSMD model resulted in comparable outcomes with respect to the aggregated model, while this was initially not the case when installations had (highly) different usage rates. In this light, it is necessary that high-resolution parameters are initialized with different distributions around some sort of measure that represents the aggregated parameter. Naturally, the aggregation of different high-resolution states must always lead to the same low-resolution state.

All in all, in order to facilitate a more thorough cross-validation process, the aggregated model should (a) capture the outcomes of interest by extracting a number of interesting exemplars from behavioural clusters by means of exploratory analysis, and (b) apply disaggregation functions in such a way that different distributions of high-resolution inputs can be created. In this way, the aggregated model can validate the high-resolution model with massive exploration while still accounting for a possible information loss due to aggregated heterogeneities in the high-resolution model.

6.2. The importance of cross-validation tests

This study did illustrate the impact of cross-validation tests with regard to the fitness for purpose of a low-resolution model. As stated in section 2.5, this study did specify a test for the error in behavioural patterns and a test for the numerical error. Without the numerical error test, it would not be hard to make the WSMD model and the aggregated model consistent as the structure of the WSMD model only involves a limited number of fundamental behavioural patterns (as illustrated by the base ensemble of the aggregated model). However, with a numerical error test, it appeared to be much harder to achieve consistency. The inclusion of a numerical test made sense for the case in this study, as a small deviation in the sailing availability could have resulted in different choices regarding the design of a force support for the vessel in question.

However, other applications can involve the specification of different cross-validation tests. For instance, if the scope would be broadened by aggregating individual military objects to the force of an entire nation or even a certain alliance, the focus may shift from numerical consistency to only behavioural consistency. Decision-makers can then explore interesting dynamics with such aggregated models, and use these insights to shape the system of interest. Of course, in most cases such a model will not accurately reflect the actual outcomes of the system of interest. However, as long as the aggregated model can sufficiently achieve its intended fitness for purpose (which is specified by the cross-validation tests), it can be of perfect use. For example, Hughes (1995) presents an extremely simple model of the dynamics in missile combat between warships, which did prove to have significant value for defence decision-makers (Lucas & McGunnigle, 2003). To provide an example for this study, the absence of other obsolescence types in the aggregated model would be unnecessary when the numerical consistency test was neglected. In that case, decision-makers could get insight in the failure pattern on a vessel level (although this is, given the context of this study, ‘begging the question’ as decision-makers eventually want insight in the order of the effect of insolvable failures).

All in all, the presence of cross-validation tests puts the usefulness of models at different resolutions in perspective and allows decision-makers to state the merits of different models more precisely. By the use of cross-validation tests, one gets insight in the comparability between models which is important for the confidence of decision-makers in the use of models at different resolutions. In contrast, separately treating models at different resolutions can obscure the merit of models as it is unclear in what cases models are useful. This may actually be one of the reasons why many organizations tend to fall back on heavily detailed models; they rather prefer a highly detailed model that accurately represents the real-world system than using an aggregate model of which it is unclear when it is useful. In this light, additional studies are needed that put effort in the evaluation of different cross-validation tests. By getting insight in the relation between different types of cross-validation tests and their implications for the comparability of models at different resolutions, MRM can be more easily applied in industrial applications.

6.3. MRM in industrial applications

However, continuing on the previous section, a breakthrough in the use of MRM by the industry also requires more research on strong consistency (i.e. achieving consistency with disaggregation functions from the final state of an aggregate model to the final state of the high-resolution model). Although high-level insights can be relevant for the adoption of broad strategies, real-world decision-makers highly value the ‘operational level’ (Davis & Bigelow, 1998). One of the main reasons for this is that real-world policies usually require a low-level representation for the evaluation of their effectiveness. However, achieving strong consistency with proper disaggregation functions is not easy.

To illustrate, consider the following case example for the policies ‘increasing the available crew working hours’ and ‘increasing the crew productivity’. Suppose that the aggregated model has complete weak consistency (which can currently not be stated as many aggregation functions could not be evaluated). With weak consistency, the failures on a vessel level are consistent with the summed failures of all individual installations in the WSMD model. Suppose that a defence decision-maker desires insight in the relative performance of both policies on the sailing availability of a vessel. The high-resolution model is very poor in exploratory analysis, and a generalization from a few high-resolution cases is generally not a good idea (Bigelow & Davis, 2003). Now, how should this question be answered?

In practice, this question will prove to be very difficult or impossible to answer in case of weak consistency. The performance of both policies is heavily dependent on the dominance of the balancing feedback-loop ‘diminishing returns’ (explained in section 4.3) for individual installations. To illustrate, the existence of a single installation that does not allow the simultaneous treatment of different failures is decisive in the eventual performance of the policies. However, on a vessel level, limited information is available about the heterogeneity of installations with respect to their failure generation and their capacity for failure treatment. In fact, the aggregated model may only be able to approximate such information with averages, which are rarely a solid representation of low-level heterogeneity and therefore can lead to improper policy recommendations.

Of course, ways exist to better capture the heterogeneity, for instance by including additional information in the aggregated model. To illustrate, the aggregated model expresses installations in component classes, which results in a failure rate for each component class. Based on this information, one may be able to 'infer' low-level information on a higher level with exploratory analysis. However, this entire process is quite arbitrary and abstract, and in practice decision-makers often prefer clear-cut detailed engineering models (Davis & Bigelow, 1998). The disaggregation process becomes even more complex when interactions between heterogenous objects at lower levels are involved. For this case, models with an object resolution higher than the WSMD model can specify interactions on a component level, while the WSMD model and the aggregated model approximate the resulting exponential degradation behaviour with a feedback-loop between the condition and degradation of aggregated objects. Although these approximations may be fine for weak consistency, the disaggregation from a 'high-level' feedback-loop to an 'agent-based' specification is very hard.

Therefore, further studies should put effort in researching strong consistency; i.e. when and how should disaggregation functions be developed for studying real-world policy effectiveness. Especially when strong consistency is infeasible, more research is needed in how models at different resolutions can together be used to tackle a problem and provide real-world policy advice. This study did illustrate the benefits of using an aggregate model, but did not translate high-level policies back to the WSMD model. All in all, the development of new theories for the combined use of models at different resolutions is essential for its applicability in industry.

6.4. Key takeaways of discussion

- A thorough cross-validation process requires the exploration of the entire model uncertainty space. An aggregate model can better enable exploratory analysis than a detailed model, but this does involve the specification of disaggregation functions. Also, because of lower level heterogeneity, different initializations on a lower level may be needed to cross-validate a single model run on a higher level.

- The presence of cross-validation tests puts the usefulness of models at different resolutions in perspective and allows decision-makers to state the merits of different models more precisely. By getting insight in the relation between the type of cross-validation test and its implications for the comparability of models at different resolutions, MRM can be more easily applied in industrial applications.

- Although low-resolution models may be important for designing broad strategies, the evaluation of real-world policies usually requires a specification on a detailed level. In this light, insights in the achievability of strong consistency and the combined use of models at different resolutions is important when providing real-world policy advice.

7. Conclusion

This study aims to answer to following main question:

"How can multi-resolution modelling be useful in improving the robustness of a naval force support system?"

The main conclusion of this study is that multi-resolution modelling can be considered useful as the detailed WSMD model and the simplified aggregated model were found to be mutually dependent in their design, validity, and use to improve the robustness of a naval force support system

- *Leg one: design*. On the one hand, as the WSMD model involves a detailed depiction of the system of interest, it formed together with aggregation functions the basis for the design of an aggregated model. On the other hand, by means of exploratory analysis with the aggregated model, some assumptions made by the WSMD model were found to be potentially problematic when evaluating the performance of the force support system. In this way, the aggregated model also can be useful for the design of the WSMD model.

- *Leg two: validity*. As the WSMD model can better represent real-world dynamics, it served together with cross-validation tests as a basis for the establishment of the fitness for purpose of the aggregated model. In fact, treating the aggregated model in isolation would obscure its merit as it would be unclear in what cases it sufficiently reflects the dynamics of interest of a force support system. Vice versa, as also stated in the previous leg, the exploratory use of the aggregated model revealed some dubious assumptions of the WSMD model, which potentially decrease the validity of the WSMD model.

- *Leg three: use*. Unlike the WSMD model, the exploratory use of the aggregated model allows decision-makers to make strategic choices amidst uncertainty. More concrete, the aggregated model revealed patterns in the robustness of the force support system that may be hard to see through with a few high-resolution cases of the WSMD model. On the other hand, as real-world policies usually require a specification on a detailed level, the WSMD model remains important for the evaluation of the effectiveness of promising actions discovered by the aggregated model.

All in all, as a force support system is characterized by a high complexity (stressing the importance of a detailed model) and massive uncertainty (stressing the importance of an aggregate model), the adoption of a multi-resolution modelling approach was found useful. The argumentation legs (which reflect the subquestions) are described more in-depth in the following sections.

7.1. Conclusions on model design

The first leg concerns the model design phase and more specifically the aggregation practices. The following question requires an answer:

"How can the required aggregations on different dimensions of resolution in the WSMD model be implemented?"

Given the cross-validation tests, the choice for the type of aggregations depended on the degree of non-linearity and object heterogeneity in the WSMD model. In fact, the main conclusion of the design phase is that glossing over qualitatively different phenomena and objects likely results in a poor aggregation. In case that 'simple' aggregation functions were inadequate, this study 'approximated' most low-level information by (a) adopting a new modelling structure, (b) creating 'abstract' object groups, and/or (c) introducing a calibration vector. However, such approximations involve trade-offs.

- *New modelling structures.* In this study, the assignment of an equivalent age and condition on an aggregate level did lead to unacceptable consistency errors. The reason for this was that (a) installations have different degradation rates, and (b) the mapping of the condition of an object to its failure rate is non-linear. Therefore, instead of using comprehensive attribute stocks like the WSMD model does, the aggregated model involves a ‘chain’ structure which allows lower-level objects to distribute across condition stages and age cohorts. However, the implementation of a chain involves a trade-off between the computational complexity (proportional to the number of distinct phases in a chain) and the consistency error (inversely proportional to the number of distinct phases in a chain).

- *Additional ‘abstract’ groups.* In this study, installation groups were converted to composition classes which allowed the aggregated model to better capture the heterogeneity of installations in the WSMD model. However, as such abstract groups are mentally uneasy to grasp, they can decrease the interpretability of the low-resolution model. Therefore, these aggregations should always be implemented in agreement with the intended users of the low-resolution model in question.

- *Calibration vectors.* Calibration vectors can be useful when new modelling structures or additional groups in the low-resolution model still lead to unacceptable consistency errors. In this study, the condition chain required calibration as the amount of condition stages in the chain were too little. However, the use of a calibration vector heavily decreases the simplicity and the interpretability of the low-resolution model. Also, calibration vectors may not hold for different initializations (although in this study the calibration vector did hold for different asset and mission configurations in the WSMD model).

However, even with such approximations, not all low-level information may be transferrable to a higher level. In this study, it was decided to ‘lose’ information on a higher level about the obsolescence attributes of installations. Given the cross-validation tests, this information could not be sufficiently approximated on a higher level. However, as this information loss would barely hurt the fitness for purpose of the aggregated model, the loss was deemed acceptable and the cross-validation tests were maintained.

.

7.2. Conclusions on model validation

In this study, model validation was considered a two-way process as both models could have implications for each other’s validity. To examine this more in-depth, the following question requires an answer:

> *“How do the WSMD model and the aggregated model affect each other’s validity?”*

- *WSMD model → aggregated model.* The WSMD model affects the validity of the aggregated model with cross-validation tests. In this study, the cross-validation tests involved the specification of (a) the individual tests, (b) the outcome area relevant for the consistency measuring, and (c) the extent to which an error can be deemed ‘tolerable’. As the cross-validation tests define the intended comparability of models, they actually determine the fitness for purpose of the low-resolution model. Cross-validation tests may require change in case of a problematic information loss (which implies a low fitness for purpose of the low-resolution model), but this also changes the intended purpose of the low-resolution model. In this study, there was no need to change the cross-validation tests as no problematic information loss was encountered.

- *Aggregated model → WSMD model*. The aggregated model has implications for the fitness for purpose of the WSMD model due to its ability for exploratory analysis. For this study, the aggregated model produced a base ensemble of runs, thereby able to validate ‘hidden’ assumptions in the WSMD model. For instance, the WSMD model implemented a fixed overhaul duration and a fixed crew productivity, thereby implicitly assuming that the overhaul crew can always manage the required amount of work in time and that installations have unlimited maintenance access. Although such relations may hold under ‘normal’ circumstances, the aggregated model did show scenarios where crews face higher workloads than expected. In such cases, balancing ‘limits to growth’ feedbacks may take over which can result in working delays. If such feedback-loops are not included, defence decision-makers may push the system in the wrong direction when trying to eliminate a performance gap.

7.3. Conclusions on model use

The development of an aggregated model next to the WSMD model was deemed important as it (a) allows reasoning on a higher level, and (b) facilitates decision-makers to make choices amidst uncertainty by means of exploratory analysis. To examine the contribution of the aggregated model more in-depth, the following question requires an answer:

“How can the aggregated model be used to increase the performance of the Dutch force support system as represented by the WSMD model?”

The main contribution of the aggregated model in its use denoted its ability to reveal patterns in the robustness of a force support system that may be hard to see through with only a few high-resolution cases of the WSMD model. Given figure 7.1, some helpful insights of robustness curves are pointwise stated.

- Robustness curves are helpful to get insight in strategic questions. For instance, a helpful insight in this study was that a linear increase in the time between two consequent overhauls (i.e. the overhaul interval) exponentially decreased the robustness of the maintenance plans as depicted by the models. In this way, decision-makers are aware that a linear increase in the overhaul interval can only be facilitated by an exponential increase in the maintenance effort.

- Robustness curves allow decision-makers to make trade-offs regarding the performance of policies. Although the robustness analysis in this study did not involve such trade-offs, policies in general can have the same ‘average’ robustness but have different robustness curves. Such curves provide a comprehensive view of the robustness of a policy and thereby allow decision-makers to make choices amidst uncertainty.

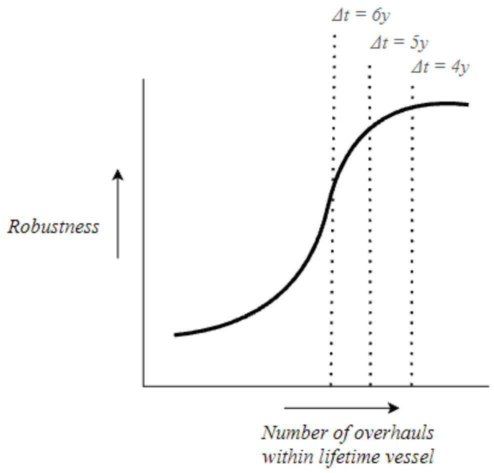

Figure 7.1. Rough representation of the robustness of a vessel against the number of overhauls in its lifetime. Δt represents the overhaul interval.

Next to robustness curves, the limited number of inputs of the aggregated model also enabled the construction of worst-case scenarios. To illustrate, scenarios characterized by a high degradation, a long overhaul interval, and/or a low crew productivity significantly decreased the performance of maintenance plans with respect to the sailing availability and the working hours. At the same time, these scenarios highlighted potential leverage points to increase the robustness of the force support system. To illustrate, an increase in the crew productivity appears to be a promising policy to increase the system robustness.

Still, the WSMD model remains crucial for the real-world implementation of such promising actions discovered by the aggregated model. As high-resolution models can better represent real-world complexity, they can better assess the actual effectiveness of policies. To illustrate, in this study limited information was available about the heterogeneity of installations on an aggregate level with respect to their failure generation and their capacity for failure treatment. However, this information is decisive for the evaluation of policies that increase the daily available working hours on a vessel. In other words, when low-resolution models are not able to disaggregate an aggregate state, the best option for low-resolution models is to identify ‘leverage regions’ that require further investigation by the high-resolution model.

7.4. Limitations

- The aggregated model could not be fully cross-validated with the WSMD model, which hindered the full establishment of the fitness for purpose of the aggregated model. This aggregated model could not be fully cross-validated for two reasons.

    1. The main reason concerns the on-going developments in the WSMD model, which hindered the cross-validation of the recovery submodel. The recovery submodel is responsible for many important dynamics in a force support system, and the inability to cross-validate this

submodel is the prime reason for why a policy ranking based on the outcomes of the aggregated model was deemed unwise.

2. This study only cross-validated the aggregated model on a few configurations of the WSMD model with a limited number of scenarios. A better cross-validation process will involve disaggregation practices, which allows models to compare for all relevant behavioural modes in the outcome space.

- Regarding model development, some (potentially) important structures are not included in the aggregated model.

  1. As appeared during the validation phase, the design of a HR system and including a balancing feedback-loop to represent diminishing returns in the employment levels of the crew are important for a proper evaluation of policies. As these structures were not included, the effectiveness of policies had to be stated with great care.
  2. At the beginning of this study, a variety of (primarily static) structures in the WSMD model were not included in the scope, of which the costs structure is deemed most important. The absence of a costs structure in the aggregated model makes it difficult to compare policies on their costs (which is in the aggregated model only represented by the maintenance hours and the renewal hours).

- The specification of policies in this study is somewhat arbitrary as this was done manually. A better evaluation will involve the use of robust optimization frameworks (Bartholomew & Kwakkel, 2020). The absence of a more sophisticated policy evaluation procedure made it difficult to state a performance ranking of policies.

7.5. Scientific research recommendations

There is ample room for further research in the MRM field. In line with the discussion in chapter 6, there is a need for more research on cross-validation and strong consistency. The final recommendation concerns the composability of models at different resolutions, which is acknowledged as one of the main open research problems in the modelling & simulation field (Castro, 2019).

Strong consistency

Regarding the type of cross-validation, this study focussed on weak consistency. However, weak consistency is generally a well-researched topic in the MRM field, while ‘strong consistency’ remains vaguely addressed. To recap, strong consistency involves the disaggregation from the final state of an aggregate model to the final state of a detailed model. The limited research towards strong consistency is surprising, as the industry usually has to specify policies on a detailed level. Without strong consistency, it is very difficult or impossible to evaluate the real-world effectiveness of policies with an aggregate model only. Therefore, more research is needed towards the feasibility of strong consistency in MRM. This inherently involves research on the development of disaggregation functions.

To continue, in cases that strong consistency is deemed infeasible, more research is needed in the combined use of models at different resolutions to tackle a problem and provide real-world policy advice. This study did illustrate that an aggregate model can identify potential leverage points in increasing a system’s robustness, but did not translate such leverage points to detailed policies that form an input to a high-resolution model. Therefore, more practical case studies are needed that illustrate how and when to go into greater detail, and when the aggregate level is sufficient for real-world policy design. Without such research, the merit of aggregate model will remain vague to practitioners, who will in turn feel more comfortable with detailed models which aim to accurately depict the real world.

Measuring consistency

In this study, the measurement of consistency involved the aggregation of the initial state of the high-resolution model to the initial state of the low-resolution model. However, more practical case studies are needed that treat the aggregate model as a starting point in the cross-validation process. In fact, as a low-resolution model can better enable exploratory analysis than a detailed model, it can identify a number of relevant behavioural modes that require consistency with the detailed model. This ‘reverse’ of the traditional cross-validation process facilitates a more thorough cross-validation process, as it allows a search in the outcome space. Note that this will inherently involve the specification of disaggregation functions from the initial state of the low-resolution model to the initial state of the high-resolution model.

Next, more research is needed in the relation between cross-validation tests and the comparability of models at different resolutions. The presence of cross-validation tests puts the usefulness of models at different resolutions in perspective and allows decision-makers to state the merits of different models more precisely. By researching the relation between a variety of cross-validation tests and the implications for the comparability of models at different resolutions, MRM becomes more understandable and attractive to the industry.

Composability

Challenges remain around the composability of MRM (Davis & Tolk, 2007; Castro, 2019). Composability refers to the ability to select and assemble model components in various combinations to satisfy specific user requirements in different contexts (National Research Council, 2006). In short, there is a strong need for conceptual and computational methods to allow an efficient lumping and un-lumping of *generic* model components with different levels of resolution (Castro, 2019). In this way, practitioners can easily choose the relevant resolutions for addressing a variety of questions in the same area of interest. In this study, the aggregated model and the WSMD model cannot be considered composable as the models were designed for a specific problem for a specific organization. In this way, the aggregated model and the WSMD model are likely to work only in a naval context, but not for assets associated to the air force or land force. Although this research did not intend to develop composable model components, it is highly valuable to research the composability of MRM in the future.

Appendix A. Validation tests

This appendix further elaborates on the structure-verification test, the parameter-verification test, the dimensional consistency test and the extreme-conditions test. The outcomes of other validation tests are covered in the main text in section 4.3.

Structure-verification & parameter-verification test

The structure-verification test compares the structure of the aggregated model with the structure of the 'modelled' real system. The parameter verification test is about the conceptual and numerical comparison of model parameters to knowledge of the real system. Conceptually, parameters should accurately depict the elements of the system structure. Numerical comparison is about the plausibility of the parameter values or uncertainty ranges. Both tests are passed when the aggregated does not object knowledge about the real system.

In general, the structure and conceptual parameters of the aggregated model are based on the structure of the WSMD model. As the structure and parameters of the WSMD model are primarily based on settled theory in the reliability engineering field, it was evaluated whether new structures and parameters in the aggregated model did not object knowledge of the modelled real system. Concrete, the new maintenance structures (the CBM structure and the maintenance priority structure) were implemented in cooperation with the model developers of the WSMD model in order to understand the dynamics of the Dutch naval force support system. In this light, the structures were considered plausible given the real system structure.

The numerical implementation of the uncertainty parameters required validation as the aggregated model was used to provide an ensemble of runs (which was not possible with the WSMD model due to its consolidative use). The bandwidths of the uncertainty parameters were also implemented in cooperation with the model developers of the WSMD model, as the bandwidths were case specific. Note that the bandwidths not only represent epistemic uncertainty but also reflect the variability in parameters due to heterogeneity of missions and installations in the WSMD model.

Dimensional consistency test

The dimensional consistency test checks for conflicting dimensions of model parameters. Fortunately, Vensim software comes with an in-built 'unit check' method. After doing the test, no unit errors were found. Six warnings were returned, but these all relate to the mapping of graphical functions. These warning can be ignored, as the graphical functions form the boundary of the modelled system and thereby allow a mapping involving different units.

Extreme-conditions test

The extreme-conditions test is performed to evaluate whether conditions can be found where to model breaks. From each submodel, a number of parameters are stress tested. The following checks have been done:

- Time between two consequent overhauls: 0. Behaviour: the operations profile fully consists of the overhaul operation. Test passed: yes.
- Number of stops per year for intermediate maintenance: 0. Behaviour: no intermediate maintenance operations (even it was scheduled beforehand). Test passed: yes.
- Allowed weeks per year intermediate maintenance: 0. Behaviour: no intermediate maintenance operations (even it stops were planned beforehand). Test passed: yes.
- Condition limitation due to age: 0. Behaviour: the condition of installation groups can always reach their maximum value. Test passed: yes.
- Tear and wear factor: 0. Behaviour: no degradation. Test passed: yes.
- Usage rate: 0. Behaviour: lower degradation, no failures. Test passed: yes.
- Adult reference MTBF[component class]: 0. Behaviour: no failures. Test passed: yes – although it would be more realistic to have very high failures (actually infinite). In this light, the effect of the condition on the MTBF is bounded until 0.0001.
- Tolerated performance: 0. Behaviour: much higher ship inavailability. the operations profile fully consists of the overhaul operation. Test passed: yes.
- Maintenance ability: 0. Behaviour: no corrective maintenance. Test passed: yes
- Crew working capacity: 0. Behaviour: no maintenance. Test passed: yes.
- Manhour norm for condition improvement: 0. Behaviour: no preventive maintenance: Test passed: yes.

Appendix B. Exploratory analysis details

This appendix contains a more precise description of the uncertainties and policies, and also contains a visualization of PRIM that is described but not presented in chapter 5.

B.1. Experimental design

The base ensemble used for validation and exploratory analysis contained 5000 scenarios. The uncertainties are described in table B.1. The policies in table B.2. are simulated for an overhaul interval of 4 years, 5 years, and 6 years. The robustness of each policy is evaluated based on 500 scenarios.

Table B.1. Specification of uncertainties for exploratory analysis

Uncertainty	Description	Range
Usage rate exploratory parameter	Parameter that varies between the minimum and maximum usage rate of a installation group during an operation.	0 to 1
Impact exploratory parameter	Parameter that accounts for uncertainty in the external impact and the usage impact on component classes during an operation	Bandwidth of 25%
Operational harshness	Parameter that varies between the minimum and maximum external impact on component classes during an operation	0 to 1
Tear and wear factor	Parameter that represents the degradation of objects due to the passing of calendar time	Bandwidth of 25%
Manhour norm for condition improvement	Parameter that represents the required manhours for a condition improvement of one condition unit	Bandwidth of 50%
Condition reduction due to age uncertainty factor	Parameter that accounts for uncertainty in the condition constraint due to age	Bandwidth of 25%
Adult reference MTBF exploratory parameter	Parameter that accounts for uncertainty in the adult MTBF of optimally performing component classes when fully used	Bandwidth of 25%
Infant MTBF exploratory parameter	Parameter that represents uncertainty in the infant MTBF of optimally performing component classes when fully used	Bandwidth of 50%
Effect condition parameter M *	Parameter that can vary the graphical failure functions according to the disruption function of Eker et al. (2014)	-0.3 to 0.8
Effect failures on performance	Parameter that represents the decrease in performance of objects after a failure	Bandwidth of 25%
Average working time to solve a failure	-	Bandwidth of 50%
Maintenance ability exploratory parameter	Parameter that varies between the minimum and maximum maintenance ability of each echelon during an operation	0 to 1
Required performance exploratory parameter	Parameter that varies between the minimum and maximum required performance of component classes during an operation	0 to 1
Average working time for an installation renewal	-	Bandwidth of 25%
Allowed weeks per year intermediate maintenance	-	5 to 20 weeks
Number of stops per year for intermediate maintenance **	-	2 to 6 stops
Time between two consequent overhauls **	-	4 to 6 years
Switch planned vs unplanned maintenance **	Switch that activates the initial maintenance plans in the WSMD model or CBM (only in the aggregated model)	0 or 1

** Please note figure B.1. for a visual representation of the uncertainty in the graphical function that represent the effect of the condition on the adult reference MTBF.*

*** Only included in the base ensemble for validation and vulnerability analysis. These 'uncertainties' are actually levers, and therefore not included as uncertainties in the robustness analysis.*

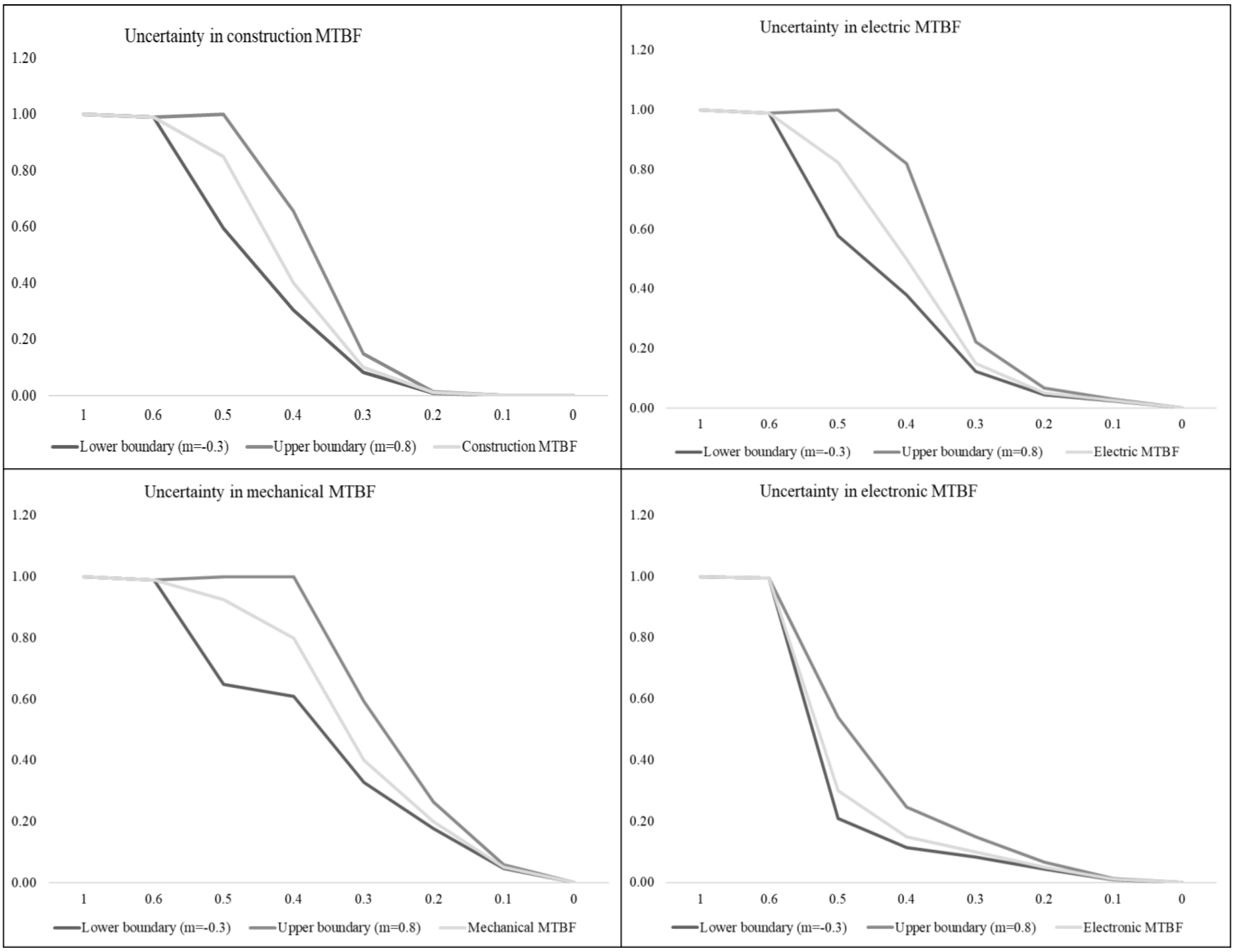

Figure B.1. Uncertainty in the effect of the condition on the adult reference MTBF of the component classes. The x-axis is specified as 1-condition[component class], and the y-axis is specified as 1-effect[component class]. For the function, please read Eker et al. (2014)

Table B.2. Specification of policies for robustness analysis. 3 iterations are performed: for an overhaul interval of 4 years, 5 years, and 6 years. Note that the values are somewhat arbitrary, a better approach would involve robustness optimization procedures.

Policy	Description	Value
ILM assistance during sailing	ILM is a crew that is initially only active during intermediate maintenance. This policy assign a part of this crew to the ‘sailing’ crew OLM.	10% of the ILM crew
Available working hours increase	Increase of the available working hours, which implies extra recruitment of personnel (the delay is not taken into account).	25% increase
Switch planned vs unplanned maintenance (2 policies)	Switch that activates the initial maintenance plans in the WSMD model or CBM (only in the aggregated model).	0 or 1
Redundancy policy	Redundancy in an object increases the delay of a functional failure (which is different from an ‘ordinary’ failure). In case of a functional failure, the ship becomes unavailable.	25% increase
Number of stops per year for intermediate maintenance	-	6. Initially: 3.
Switch condition threshold	Objects below this threshold are renewed during overhaul. This policy increases this threshold.	All objects below condition 0.6 are renewed. Initially: 0.5.
Productivity increase	Due to training or improved object design, the productivity of the crews can increase.	75% of initially required manhours/failure and manhours/condition unit for corrective maintenance and preventive maintenance respectively.

B.2. PRIM

Figure B.2. shows parameters that significantly contribute to a sailing availability below 95%. A sailing availability below 95% is considered a worst case. To interpretate the results, consider the coverage and density at the top right. To clarify, the restricted parameters are found significant in contributing to a worst-case outcome region where about 78% of all scenarios were considered worst-case. Furthermore, this region contained 36% of all worst cases in the entire uncertainty space. Achieving a higher coverage would lead to undesired low density values, therefore this worst-case region is chosen. The restricted bandwidths of the parameters are found statistically significant as their p values are below 0.025 (i.e. a one-sided p-value). The p-values of the parameters are presented in brackets.

Now, the question is whether these statistics allow the construction of scenarios. The density value demonstrates that most scenarios (i.e. almost 4 out of 5) are worst-case for the given bandwidths of these parameters. As a density of 80% is usually taken as a rule of thumb for scenario construction (Bryant & Lempert, 2010), the restricted parameters can be interpreted as the scenarios stated in chapter 5. However, it should be noted that these restricted parameters only capture 36% of the total number of worst-case scenarios. In this light, it should be kept in mind that the set of restricted parameters is not exclusive for scenario construction.

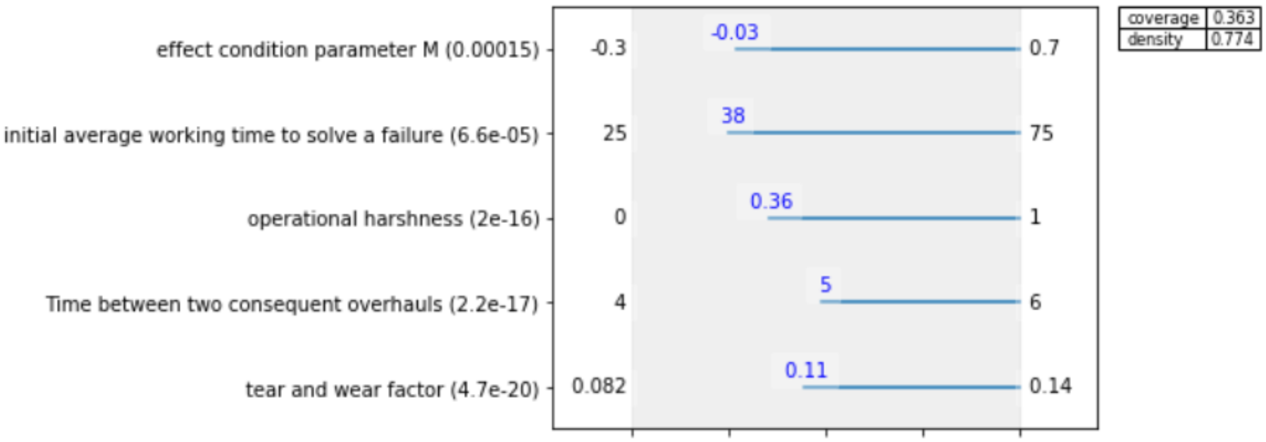

Figure B.2. PRIM outcome visualized for a sailing availability below 95%. Please read section 2.2. for an explanation of PRIM

References

Abbass, H. A., Bender, A., Dam, H. H., Baker, S., Whitacre, J., & Sarker, R. (2008). Computational scenario-based capability planning. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation* (pp. 1437-1444). Australian Defence Force Academy.

Accenture. (2017). Defining standards with service-level agreements (No. 172835). https://www.accenture.com/t20170814T082829Z__w__/us-en/_acnmedia/Accenture/Designlogic/17-2860/documents/Accenture-Defining-Standards-with-Service-Level-Agreements.pdf

Adamides, E. D., Stamboulis, Y. A., & Varelis, A. G. (2004). Model-based assessment of military aircraft engine maintenance systems. In *Journal of the Operational Research Society*, 55(9), 957-967.

Ahram, T. Z., Karwowski, W., Sala-Diakanda, S., & Jiang, H. (2017). Modeling Decision Flow Dynamics for the Reliable Assessment of Human Performance, Crew Size and Total Ownership Cost. In *Advances in Applied Digital Human Modeling and Simulation* (pp. 117-129). Springer, Cham.

Auping, W. L., Pruyt, E., & Kwakkel, J. H. (2015). Societal ageing in the Netherlands: a robust system dynamics approach. In *Systems Research and Behavioral Science*, 32(4), 485-501.

Auping, W. (2018). Synthesis. In *Modelling uncertainty: Developing and using simulation models for exploring the consequences of deep uncertainty in complex problems* (doctoral dissertation). Delft University of Technology, Delft.

Bankes, S. C. (1993). Exploratory Modeling for Policy Analysis. In *Operations Research*, 4 (3), 435-449.

Bankes, S. C., Walker, W. E., & Kwakkel, J. H. (2013). Exploratory modeling and analysis. *Encyclopedia of operations research and management science,* 532-537.

Bartholomew, E., & Kwakkel, J. H. (2020). On considering robustness in the search phase of robust decision making: a comparison of many-objective robust decision making, multi-scenario many-objective robust decision making, and many objective robust optimization. In *Environmental Modelling & Software*, 127, 104699.

Bender, A., Pincombe, A. H., & Sherman, G. D. (2009). Effects of decay uncertainty in the prediction of life-cycle costing for large scale military capability projects. In *18th World IMACS Congress and MODSIM 2009 - International Congress on Modelling and Simulation: Interfacing Modelling and Simulation with Mathematical and Computational Sciences, Proceedings* (pp. 1573–1579). Modelling and Simulation Society of Australia and New Zealand Inc. MSSANZ.

Betts, R. K. (1982). Surprise attack: Lessons for defense planning. Washington, D.C: Brookings Institution.

Bigelow, J. H., & Davis, P. K. (2003). Implications for model validation of multiresolution, multiperspective modeling (MRMPM) and exploratory analysis. Santa Monica, CA: RAND Corporation.

Bowers, J., Elsawah, S., & Ryan, M. (2017). Reusable modules to support rapid model building: A case study of defence force design. In *INCOSE International Symposium*, 27(1), 1539-1553.

Bryant, B. P., & Lempert, R. J. (2010). Thinking Inside the Box: a participatory computer-assisted approach to scenario discovery. In *Technological Forecasting and Social Change*, 77(1), 34-49.

Burk, R. C., & Parnell, G. S. (2011). Portfolio decision analysis: Lessons from military applications. In *Portfolio decision analysis (pp. 333-357).* Springer, New York, NY.

Carchia, M. (1999). Electronic/Electrical Reliability. Retrieved from https://users.ece.cmu.edu/%7Ekoopman/des_s99/electronic_electrical/

Caron, J. D., Fong, V., & Brion, V. (2019). On the Use of Simulation and Optimization for Mission Modules Selection in a Maritime Context. In *Military Operations Research*, 24(1), 41-56. JSTOR

Castro, R. (2019). Open research problems. In Zeigler, B. P., Muzy, A., and Kofman, E. (Eds)., *Theory of modelling and simulation: discrete event iterative system computational foundations*, pp. 641-658. Elsevier Academic Press.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. In *Geoscientific model development*, 7(3), 1247-1250.

Checco, J., Berg, B., & Loerch, A. (2017). Optimizing US Army Force Size Under Uncertainty Through Stochastic Programming. In *Military Operations Research,* 22(2), 19-38. JSTOR

Clark, R. H., & Pisani, A. A. (1985). Defense Resource Dynamics. In *the Proceedings of the 1985 International System Dynamics Conference.* International System Dynamics Society.

Coyle, R. G., & Gardiner, P. A. (1991). A system dynamics model of submarine operations and maintenance schedules. In *Journal of the Operational Research Society, 42(6), 453-462.*

Coyle, R. G. (1996). System dynamics applied to Defense analysis: a literature survey. *Def Anal 12(2):141–160.*

Coyle, R. G., Exelby, D. & Holt, I. (1999). System dynamics in defence analysis: some case studies. *J Oper Res Soc 50:372–382.*

Davis, P. K., & Finch, L. (1993). Defense planning for the Post-Cold War Era. *Giving Meaning to Flexibility, Adaptiveness, and Robustness of Capability.* Santa Monica, CA: RAND Corporation.

Davis, P. K., & Hillestad, R. (1993). Families of models that cross levels of resolution: Issues for design, calibration and management. In *Proceedings of 1993 Winter Simulation Conference-(WSC'93)* (pp. 1003-1012). IEEE.

Davis, P. K. (1995). Aggregation, Disaggregation, and the 3:1 Rule in Ground Combat. Santa Monica, CA: RAND Corporation.

Davis, P. K., & Bigelow, J. H. (1998). Experiments in multiresolution modeling (MRM). Santa Monica, CA: RAND Corporation.

Davis, P. K. (2003). Exploratory analysis and implications for modeling. RAND-PUBLICATIONS-MR-ALL SERIES-, 255-284.

Davis, P. K., & Tolk, A. (2007). Observations on new developments in composability and multi-resolution modeling. In *2007 Winter Simulation Conference* (pp. 859-870). IEEE.

Davis, P. K. (2014). Analysis to inform defense planning despite austerity. Santa Monica, CA: RAND Corporation.

Davis, P. K. (2016). Capabilities for joint analysis in the department of defense: rethinking support for strategic analysis. Santa Monica, CA: RAND Corporation.

Davis, P. K. (2018). Defense planning when major changes are needed. In *Defence studies, 18(3), 374-390*. Taylor & Francis.

Defense Modeling and Simulation Enterprise (MSE). (n.d.). M&S Glossary. Retrieved from https://www.msco.mil/MSReferences/Glossary/TermsDefinitionsE-H.aspx

De Spiegeleire, S. (2011). Ten trends in capability planning for defence and security. *The RUSI Journal, 156(5), 20-28.*

Dhillon, B. S. (2006). Introduction to engineering reliability. In *Maintainability, Maintenance, and Reliability for Engineers*. CRC press.

Dutch Ministry of Defence. (2020a). *Defensievisie 2035: Vechten voor een veilige toekomst.* Retrieved from https://www.defensie.nl/onderwerpen/defensievisie-2035

Dutch Ministry of Defence. (2020b). *Zr.Ms. Friesland.* Retrieved from https://www.defensie.nl/organisatie/marine/eenheden/schepen/zr.-ms.-friesland

Eisler, C., & Allen, D. (2012). A Strategic Simulation Tool for Capability-Based Joint Force Structure Analysis. In *Proceedings of the International Conference on Operations Research and Enterprise Systems, 21-30, 4-6.* Defence R&D Canada.

Eker, S., Slinger, J., van Daalen, E., & Yücel, G. (2014). Sensitivity analysis of graphical functions. In *System Dynamics Review*, 30(3), 186-205.

Elsawah, S., Ryan, M. J., Gordon, L., & Harris, R. (2018). Model-based Assessment of the Submarine Support System. In *INCOSE International Symposium (Vol. 28, No. 1, pp. 392-406).*

Esmaeili, E., Karimian, H., & Bisheh, M. N. (2019). Analyzing the productivity of maintenance systems using system dynamics modeling method. In *International Journal of System Assurance Engineering and Management*, 10(2), 201-211.

Fan, C. Y., Fan, P. S., & Chang, P. C. (2010). A system dynamics modeling approach for a military weapon maintenance supply system. In *International Journal of Production Economics*, 128(2), 457-469.

Fitzsimmons, M. (2006). The problem of uncertainty in strategic planning. In *Survival*, 48(4), 131-146. Taylor & Francis.

Forrester, J.W. (1961). Industrial Dynamics. Cambridge, MA: The MIT Press. Reprinted by Pegasus Communications, Waltham, MA.

Forrester, J. W. (1973). Counterintuitive behaviour of Social Systems. *Towards Global Equilibrium, eds.* D. L. Meadows and D. H. Meadows. Cambridge, MA: Wright-Allen Press.

Friedman, J. H., Fisher, N. I. (1999). Bump hunting in high-dimensional data. In *Statistical Computing* 9(2), 123–143.

Gallagher, M. A., Caswell, D. J., Hanlon, B., & Hill, J. M. (2014). Rethinking the hierarchy of analytic models and simulations for conflicts. In *Military Operations Research*, 19(4), 15-24. JSTOR.

Gorjian, N., Ma, L., Mittinty, M., Yarlagadda, P., & Sun, Y. (2010). A review on degradation models in reliability analysis. In *Engineering asset lifecycle management (pp. 369-384)*. Springer, London.

Gray, C. S. (2008). Coping with uncertainty: Dilemmas of defense planning. In *Comparative Strategy*, 27(4), 324-331.

Gray, C. S. (2014). Introduction: Defence Planning - a Mission about Consequences. In *Strategy and defence planning: meeting the challenge of uncertainty* (pp. 1-16*)*. Oxford University Press, USA.

Hamarat, C., Kwakkel, J. H., & Pruyt, E. (2013). Adaptive robust design under deep uncertainty. In *Technological Forecasting and Social Change*, 80(3), 408-418.

Hamarat, C., Kwakkel, J. H., Pruyt, E., & Loonen, E. T. (2014). An exploratory approach for adaptive policymaking by using multi-objective robust optimization. In *Simulation Modelling Practice and Theory*, 46, 25–39.

Harrison, K. R., Elsayed, S., Garanovich, I., Weir, T., Galister, M., Boswell, S., … Sarker, R. (2020). Portfolio Optimization for Defence Applications. In *IEEE Access*, 8, 60152–60178. IEEE.

Hastings, N. A. J. (2015). Reliability, Availability, and Maintainability. In *Physical Asset Management* (pp. 373-411). Springer, Cham.

Helton, J. C., & Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. In *Reliability Engineering & System Safety*, 81(1), 23-69.

Hughes Jr, W. P. (1995). A salvo model of warships in missile combat used to evaluate their staying power. In *Naval Research Logistics (NRL),* 42(2), 267-289.

Karandaev, A. S., Khramshin, V. R., Evdokimov, S. A., Kondrashova, Y. N., & Karandaeva, O. I. (2014). Metodology of calculation of the reliability indexes and life time of the electric and mechnical systems. In *2014 International Conference on Mechanical Engineering, Automation and Control Systems (MEACS)* (pp. 1-6). IEEE.

Kwakkel, J. H. (2017). The Exploratory Modeling Workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. *Environmental Modelling & Software, 96, 239-250.*

Lane, D.C. (2000). Diagramming conventions in System Dynamics. In *Journal of the Operational Research Society* 51(2), 241–245.

Lempert, R. J., Popper S. W., Bankes S. C. (2003). Shaping the next one hundred years: new methods for quantitative, long-term policy analysis, *MR-1626-RPC*. RAND, Santa Monica

Loper, M. L., & Register, A. (2015). Introduction to modeling and simulation. In *Modeling and Simulation in the Systems Engineering Life Cycle (pp. 3-16)*. Springer, London.

Lucas, T. W., & McGunnigle, J. E. (2003). When is model complexity too much? Illustrating the benefits of simple models with Hughes' salvo equations. In *Naval Research Logistics (NRL)*, 50(3), 197-217.

Ma, F. (2019). Exploratory dynamic capacity analysis of defense forces. In *Proceedings of the 2019 Summer Simulation Conference (pp. 1-12)*. SCS

Maier, H. R., Guillaume, J. H. A., van Delden, H., Riddell, G. A., Haasnoot, M., & Kwakkel, J. H. (2016). An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? *Environmental Modelling & Software*, 81, 154-164.

Malmi, E., Pettersson, V., Syrjänen, S., Nissinen, N., Åkesson, B., & Lappi, E. (2011). Warfare simulation and technology forecasting in support of military decision making. In *INFOCOMP*, 23-29.

Marlow, D., & Novak, A. (2013). *Fleet Sizing Analysis Methodologies for the Royal Australian Navy's Combat Helicopter Replacement Project.* Defence Science and Technology Organisation Fishermans Bend (Australia) Joint and Operations Analysis Div.

McLucas, A., Lyell, D., & Rose, B. (2006). Defence capability management: Introduction into service of multi-role helicopters. In *Proceedings of the 24th International Conference of the System Dynamics Society* (pp. 92-110).

McLucas, A.C. (2011). Using system dynamics modelling to aid in establishing realistic availability for complex systems. In *the proceedings of systems engineering test and evaluation conference*. Adelaide.

McLucas, A. C., & Elsawah, S. (2020). System Dynamics Modeling to Inform Defense Strategic Decision-Making. *In System Dynamics.* pp. 341–373. Springer US.

McPhail, C., Maier, H. R., Kwakkel, J. H., Giuliani, M., Castelletti, A., & Westra, S. (2018). Robustness metrics: How are they calculated, when should they be used and why do they give different results? In *Earth's Future*, 6(2), 169-191.

Moallemi, E. A., Elsawah, S., Turan, H. H., & Ryan, M. J. (2018). Multi-objective decision making in multi-period acquisition planning under deep uncertainty. In *2018 Winter Simulation Conference (WSC)* (pp. 1334-1345). IEEE.

Moallemi, E. A., Kwakkel, J., de Haan, F. J., & Bryan, B. A. (2020). Exploratory modeling for analyzing coupled human-natural systems under uncertainty. *Global Environmental Change*, 65, 102-186. Elsevier.

Morecroft, J. D. W. (1982). A critical review of diagramming tools for conceptualizing feedback system models. In *Dynamica* 8, 20–29.

National Research Council (NRC). (2006). Defense modeling, simulation, and analysis: meeting the challenge. In *Committee on Modeling and Simulation for Defense Transformation*. National Academic Press.

NATO. (2018). *Analysis Support Guide for Risk-Based Strategic Planning.* Retrieved from https://www.centrostudiesercito.it/doc/NATO_Risk_Based_Strategic_Planning.pdf

NATO. (2020). *NATO 2030: United for a New Era*. Retrieved from https://www.nato.int/cps/en/natohq/news-_179840.htm

Navarro-Galera, A., Ortúzar-Maturana, R. I., & Muñoz-Leiva, F. (2011). The application of life cycle costing in evaluating military investments: An empirical study at an international scale. In *Defence and Peace Economics*, 22(5), 509-543.

Oxenham, D. (2010). The next great challenges in systems thinking: A defence perspective. *Civil Engineering and Environmental Systems*, 27(3), 231–241.

Rabelo, L., Kim, K., Park, T. W., Pastrana, J., Marin, M., Lee, G., ... & Gutierrez, E. (2015). Multi resolution modeling. In *2015 Winter Simulation Conference (WSC)* (pp. 2523-2534). IEEE.

Rashedi, R., & Hegazy, T. (2016). Holistic analysis of infrastructure deterioration and rehabilitation using system dynamics. In *Journal of Infrastructure Systems,* 22(1), 1-10.

Schneller, G. O., & Sphicas, G. P. (1983). Decision making under uncertainty: Starr’s domain criterion. In *Theory and Decision*, 15(4), 321–336.

Senge, P. M., & Forrester, J. W. (1980). Tests for building confidence in system dynamics models. In *System dynamics, TIMS studies in management sciences,* 14, 209-228.

Shafi, K., Elsayed, S., Sarker, R., & Ryan, M. (2017). Scenario-based multi-period program optimization for capability-based planning using evolutionary algorithms. In *Applied Soft Computing*, 56, 717-729. Elsevier Academic Press

Sharma, P., Kulkarni, M. S., & Yadav, V. (2017). A simulation based optimization approach for spare parts forecasting and selective maintenance. In *Reliability Engineering & System Safety*, 168, 274-289.

Shephard, R. W., & Färe, R. (1974). The law of diminishing returns. In *Production theory* (pp. 287-318). Springer, Berlin, Heidelberg.

Steinmann, P., Auping, W. L., & Kwakkel, J. H. (2020). Behavior-based scenario discovery using time series clustering. In *Technological Forecasting and Social Change*, 156, 120052.

Sterman, J. D. (1988). A skeptic's guide to computer models. *Foresight and National Decisions, ed. L. Grant, pp. 133-169.* Lanham, MD: University Press of America.

Sterman, J. D. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw, New York.

Suzuki, S., Stern, D., & Manzocchi, T. (2015). Using association rule mining and high-dimensional visualization to explore the impact of geological features on dynamic flow behaviour. In *SPE annual technical conference and exhibition.*

Tate, D. M., & Thompson, P. M. (2017). Portfolio selection challenges in defense applications. Inst. Defense Analyses, Alexandria, VA, USA, Tech. Rep. NS D-8493.

Trouw. (2021). Geef ons geld, zeggen deze topmannen van Defensie: ‘Wij zijn de klaplopers van de Navo.’ Retrieved from https://www.trouw.nl/binnenland/geef-ons-geld-zeggen-deze-topmannen-van-defensie-wij-zijn-de-klaplopers-van-de-navo~bcd8c019/

Turan, H. H., Elsawah, S., & Ryan, M. J. (2020). A long-term fleet renewal problem under uncertainty: A simulation-based optimization approach. In *Expert Systems with Applications, 145, 13-158.*

Ventana Systems. (2010). *Vensim Reference Manual*. Harvard, MA: Ventana Systems.

Walker, W.E., Marchau, V.A.W.J., Kwakkel, J.H. (2013). Uncertainty in the framework of Policy Analysis, In *Thissen, W.A.H., Walker, W.E. (Eds.), Public Policy Analysis: New Developments*. Springer, Berlin, Germany.

Wesolkowski, S., & Eisler, C. (2015). Capability-Based Models for Force Structure Computation and Evaluation. In *NATO Workshop on Integrating Modelling & Simulation in the Defence Acquisition Lifecycle and Military Training Curriculum*, 1–22. Defence R&D Canada.

Wojtaszek, D., & Wesolkowski, S. (2012). Military Fleet Mix Computation and Analysis. In *IEEE Computational Intelligence Magazine,* 7(3), 53-61.

Wolstenholme, E. (1990). System Enquiry. In *A system dynamics approach.* Chichester: John Wiley and Sons.

Xiong, J., Liu, J., Chen, Y., & Abbass, H. A. (2014). A knowledge-based evolutionary multiobjective approach for stochastic extended resource investment project scheduling problems. In *IEEE Transactions on Evolutionary Computation*, 18(5), 742-763.

Yücel, G., & Barlas, Y. (2011). Automated parameter specification in dynamic feedback models based on behavior pattern features. In *System Dynamics Review*, 27(2), 195-215.

Young, L. (2015). Defining and developing soft capabilities within defence. In *21st International Congress on Modelling and Simulation.* BMT.

Zeigler, B. P., Muzy, A., and Kofman, E. (2019). Basic formalisms: DEVS, DESS, DTSS. In *Theory of modelling and simulation: discrete event iterative system computational foundations*, pp. 153-165. Elsevier Academic Press.

Zhang, S. T., Dou, Y. J., & Zhao, Q. S. (2014). Evaluation of capability of weapon system of systems based on multi-scenario analysis. In *Advanced Materials Research* (Vol. 926, pp. 3806-3811). Trans Tech Publications Ltd.