CRANFIELD UNIVERSITY

PETROS BOUTSELIS

INVESTIGATING THE APPLICABILITY OF BAYESIAN NETWORKS TO DEMAND FORECASTING DURING THE FINAL PHASE OF SUPPORT OPERATIONS

CRANFIELD DEFENCE AND SECURITY (CDS)

PhD
Academic Year: 2014 – 2019

Supervisor: Dr Ken McNaught
Associate Supervisor: Dr Adam Zagorecki
March 2019

CRANFIELD UNIVERSITY

CRANFIELD DEFENCE AND SECURITY

PhD

Academic Year 2014 – 2019

PETROS BOUTSELIS

Investigating the Applicability of Bayesian Networks to Demand Forecasting During the Final Phase of Support Operations

Supervisor: Dr Ken McNaught
Associate Supervisor: Dr Adam Zagorecki
March 2019

This thesis is submitted in partial fulfilment of the requirements for the degree of PhD

ABSTRACT

A challenge faced by businesses that provide logistical support to systems is when the provision of those support services is no longer required. A typical example of such a situation is when military operations come to an end. In such cases, those companies that have a contract with the Armed Forces to provide maintenance support for the deployed systems, need to maintain those systems at minimum cost during that final phase, that is from the time the decision to stop the operations is announced until their very end.

During the final phase, a challenging problem is forecasting the demand for spare parts, corresponding to equipment failures within the system. This is because the support context, the number of supported systems, the support equipment or even the operational demand can change during that period, and also because there can be very limited opportunities to place orders to cover demand.

This thesis suggests that these types of problems can take advantage of the data that have been collected during the support operations prior to the initiation of the closing down process. Moreover, the thesis investigates the exploitation of these data by the use of Bayesian Networks to forecast the demand for spares that will be required for the provision of maintenance during the final phase.

The research uses stochastically simulated Support Chain scenarios to generate data and also to evaluate different methods of constructing Bayesian Networks. The simulated scenarios differ in the demand context as well as in the complexity of the Equipment Breakdown Structure of the supported systems. The Bayesian Networks' structure development methods that are tested include unsupervised machine learning, eliciting the structure from Subject Matter Experts, and two hybrid approaches that combine expert elicitation and machine learning. These models are compared to respective logistic regression models, as well as subject matter experts-adjusted single exponential smoothing forecasts.

The comparison of the models is made using both accuracy metrics and accuracy implication metrics. These forecast models' comparison methods are analysed in order to evaluate their appropriateness. The analyses have provided a number

of novel outputs. The algebraic analysis of the accuracy metrics theoretically proves empirical problems that have been discussed in the literature but also reveals others. Regarding the accuracy implication metrics, the analysis shows that for the particular type of problems examined in this thesis –final phase problems – the accuracy implication metrics commonly applied are not enough to inform decision making, and a number of additional ones are required.

The research shows that for the scenarios examined, the Bayesian Networks that had their structure learned using an unsupervised algorithm performed better in the accuracy metric than any of the other models. However, even though these Bayesian Networks also did well with the accuracy implication metrics, neither they, nor any of the others was consistently dominant. The reason for the discrepancy in the results between the accuracy and the accuracy implication metrics is that the latter are not only driven by how accurate the forecast model's prediction is, but also by the model of the residual error and the bias.

Keywords:

Bayesian Networks, operational availability, spare parts forecasting, accuracy implication metrics, support chain simulation

ACKNOWLEDGEMENTS

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my special appreciation and thanks to my supervisor Dr Ken McNaught. Ken, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for showing me inspiring ways out of my occasional deadlocks. Your advice on both research as well as on my career have been priceless.

Moreover, Dr Kevin Burgess deserves a really special thanks. Kevin, despite you had no obligation to support me, you have been tirelessly providing me with courage and invaluable advice. I have being so lucky to have a unique person like yourself on my side through this journey.

I would also like to thank Dr John Salt and Dr Adam Zagorecki for serving as my committee members. I want to thank you for letting my defence be an enjoyable moment, and for your comments and suggestions.

A special thanks to my family. Words cannot express how grateful I am to my beloved wife Chrysa. The sacrifices that you made on my behalf were key to my progress. I would also like to thank my children Stavroula (Stevi), Konstantinos-Stavros and Alexandros. My little angels, the thought that you looked upon me meant that I could look only forward.

Last but not least I would like to express my appreciation to my parents and parents-in-law. Stavro and Stavroula, your sleepless prayers and love have been a blessing to me every day of my life. Kosta and Vasiliki, one of the reasons I have been able to move forward is that I have you covering for me when I am not there.

TABLE OF CONTENTS

LIST OF FIGURES

Appendices

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AD	Activity Diagram
AE	Absolute Error
AMF	Accuracy Metric Function
APE	Absolute Percentage Error
AvgRelMAE	Average Relative Mean Absolute Error
BN	Bayesian Network
CPT	Conditional Probability Table (same as NPT)
DAG	Directed Acyclic Graph
DP	Discardable Part
DU	Disposable Unit
EBS	Equipment Breakdown Structure
EOL	End of Life
FB	Forward Base
FPP	Final Phase Problem
GMAE	Geometric Mean Absolute Error
GRMSE	Geometric Root Mean Squared Error
GS	Grow Shrink
IAMB	Incremental Association Markov Blanket
IID (or iid)	Independent and Identically Distributed
ISTAR	Intelligence, Surveillance, Target Acquisition, and Reconnaissance
LRU	Line Replaceable Unit
LSO	Logistic Support Organisation
LTB	Last-Time Buy
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MALDT	Mean Administrative and Logistic Delay Time
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MdAPE	Median Absolute Percentage Error
MIME	Multi-Indenture Multi-Echelon
MLDT	Mean Logistic Delay Time

MPA	Maritime Patrol Aircraft
MSE	Mean Squared Error
MTBF	Mean Time Between Failures
MTBM	Mean Time Between Maintenance
MTTR	Mean Time To Repair
MUF	Main Useful Function
NPT	Node Probability Table (same as CPT)
NV	Newsvendor
NVP	Newsvendor Problem
PB	Percentage Better
PBt	Percentage Best
PRU	Partly Repairable Unit
RAE	Relative Absolute Error
RGMAE	Relative Geometric Mean Absolute Error
RGRMSE	Relative Geometric Root Mean Squared Error
RMAE	Relative Mean Absolute Error
RMSE	Root Mean Squared Error
sAPE	symmetric Absolute Percentage Error
SC	Support Chain
SE	Squared Error
SES	Simple (or Single) Exponential Smoothing
sMAPE	symmetric Mean Absolute Percentage Error
sMdAPE	symmetric Median Absolute Percentage Error
SME	Subject Matter Expert
UAV	Unmanned Air Vehicle

1 INTRODUCTION

1.1 Problem Statement

A very challenging period for the support of systems that are used in military or business operations, is the one that follows the announcement of the decision to bring operations to an end. Several uncertainties are triggered from such a decision. What will the operational demand be during the planned final period? How many systems will be left to operate and how is this operating context going to affect the requirements for support? Withdrawing means modifying or even taking away equipment and support facilities and probably moving or withdrawing personnel including operators and mechanics with different levels of experience. So, will those involved in the final phase be able to cope with the support requirements? Depots and inventories are eventually going to be gradually reduced and replenishment lead-times could change. What will the effect of such possible changes be on the availability of the systems left behind to continue the operations until the very last one has left?

In order to be able to deliver the anticipated operational output during this final period, at the very least managers need to be able to decide on the levels of inventory to keep for the remaining supported systems given the planned changes. Therefore, at the final replenishment of the inventory, if the inventory managers place an order for their depots and obtain more than they would eventually need, then they will have incurred overage costs for the provisioning, holding and for the transport back of the excessive inventory that they will not have used. Furthermore, the parts, which are not economic to be returned or are characterised as security-sensitive, will probably need to be destroyed. Conversely, holding inventory levels that are below requirements will create an array of problems for those personnel left to run the system. For the purposes of this research, the finite time-horizon support problem with the above characteristics is called the Final Phase Problem (FPP), and as shown further below but in more details in Sections 2.2 and 2.3, the FPP has distinctively different characteristics than the similar problems studied in the literature of the Newsvendor (NVP) and the Last Time Buy / End of Life (LTB/EOL).

In order to estimate the optimum level of inventory, a forecast of the demand for spares is required. This forecast is the focus of the present thesis. During infinite time-horizon logistic operations, very common demand forecasting methods are variations of time series (Petropoulos, Makridakis, Assimakopoulos, and Nikolopoulos, 2014). Time series have also been suggested in cases similar to the ones examined in this research (Alwan, Xu, Yao, and Yue, 2016). However, as Dekker, Pinçe, Zuidwijk, and Jalil (2013) suggest, it would sometimes be useful in the forecast to be able to consider different attributes of the installed base, like the number of systems that are supported and their usage rate. Indeed, as shown in the examples of Section 1.2, during the final phase of operations, the usage rate might not be the same as in the phases before that (Phases II and IV in the Committee on Force Multiplying Technologies for Logistics Support to Military Operations and Board on Army Science and Technology, (2014, figs 2–1)), and also the number of systems that are supported might change. This suggests that the forecasts provided by time series alone without any consideration of the demand context during the final phase can be poor, since the forecast provided will be the same regardless of any information about the expected changes (Boutselis and McNaught, 2018). In practice, decision makers tend to adjust the forecasts provided by their models and especially when there are anticipated changes (Christopher, 2016, Chapter 5; Rekik, Glock, and Syntetos, 2017), using their experience and intuition and thus, in the final phase of operations they would be expected to do so too.

Nevertheless, there are approaches such as regression that provide forecasts of a response variable using scenario factors as predictors. In the forecast models for the FPP cases dealt with in this thesis, factors related to the number of systems operating and supported, as well as to their operational usage and the support resources have been used as predictors with the demand for spares being the response variable. As the literature shows (Chapter 2), such models have been applied in steady-state problems (see e.g. Sherbrooke, 2000) and it would be logical to use them for the FPPs of support operations as well. However, as demonstrated in the literature review (Chapter 2), regression models and indeed the kind of models that use predictors have rarely been applied to the

types of problems that are examined here, that is in the final phase of support operations. Furthermore, when they have been used, there is a lack of understanding of which are the most influential factors that should be included as predictors.

The latter observation is discussed in the literature review in Chapter 2. As it is further shown there, the examined published research in areas similar to the FPP have not dealt with their respective problems by seeing the Support Chain (SC) of the operations as a whole entity, that is as a system. Consequently, in taking that view, some exploration of the factors which influence the demand during the final phase was necessary. This involved the elicitation of relevant domain knowledge from Subject Matter Experts (SMEs) in order to complement what had already been identified in the literature (Chapter 3).

The FPP examined in this thesis has a number of similarities to single-period forecasting problems like those dealt with in the well-known Newsvendor problem (NVP) (Khouja, 1999) and in the “Last time buy” / “End of life” (LTB/EOL) problem which are discussed in Sections 2.2 and 2.3 respectively. The similarities of these three problem types (NVP, LTB/EOL and FPP) stem from the fact that the decision maker needs to use a demand forecast in order to optimise the supply order that he/she can place at the beginning of a limited time period ahead, with much uncertainty about the demand distribution due to the effects of the demand context that will follow. The dilemma is to decide on the inventory level of goods that should neither be more than needed since the excess inventory creates overage costs, nor less than needed since the demand that will not be met will create underage costs. Overage costs are defined within the literature of the NVP as the cost for any items that cannot be sold, while underage costs are defined as the cost for not meeting the demand (see e.g. Alwan et al., (2016))

The NVP, the LTB/EOL problems and the FPP occasionally share another common challenge, for the final reprovisioning of the inventory described above to come from only a single order. The reason for this additional challenge is that the time to make the decision can be very tight with little opportunity for subsequent corrective orders. The cause is that in extensive networks like those

that support overseas operations, the lead times can be quite long and consequently, only few resupply orders can usually be made before any additional data can be collected so as to gain more information and understanding on the new, final demand context.

In summary, the challenges that are shared by the three similar categories of problems – the NVP, the LTB/EOL and the FPP – are that the decision on the resupply order levels needs to be such that there is a balance between the likelihood that at the end of the single period ahead there are no leftovers and that there is no shortage. This decision cannot be easily adjusted. Consequently, these challenges call for an extensive analysis of the available data, information and knowledge existing up to the moment of the decision. However, as shown in more detail in Chapter 2 and in Chapter 3 (Sections 3.2, 3.4), there are a number of differences in the availability of data, information and knowledge between the FPPs which are of interest to this research, and the NVP or the LTB/EOL problems. In the latter two, the indirect assumption is that the decision maker has visibility only of her own node in the Support Chain, usually the supply part, and thus, as it is also demonstrated in Chapter 2 (Sections 2.2 and 2.3), this limits the accessibility and availability of the data and information that can be used, and this limitation also has an effect on the type of demand forecasting models that can be developed.

On the other hand, in the FPP cases which are of interest to this research, due to the assumed performance-based and availability-based contracts (Section 1.2), there is a strong collaboration between the service providers and the customer, and consequently, there is access to data from different levels and nodes of the Support Chain. Such datasets that can be found e.g. as recorded incidents in logbooks, have captured situations and conditions that no single operator / decision maker / expert can holistically acquire and contain. Therefore, the information that is contained in these datasets, if captured and analysed can potentially complement and advise the decision makers when facing challenges such as the ones discussed earlier, that is having to define the level of spares to keep in inventory given that there can be many changes to the operations and their support, and that the decisions will affect a single, final period.

The literature review in Chapter 2 demonstrates that similar challenges have been addressed in the past. However, the review also suggests a weakness in that neither the context nor the type and the availability of the data have been considered, and that consequently, the types of models that can use these data have not been explored either.

The work presented here is an attempt to address the decision maker's need for a better-informed model of the demand in order to improve decision making when confronted with the challenges of the FPP. Of specific interest is the exploitation of the data records of incidents and activities kept in the logbooks of the different functional and operational nodes of the Support Chain of operations during the phases prior to the final phase. Under the close relationships assumed to exist in the FPP settings due to availability and performance based contracts, these data records are able to be used in the demand forecast models.

1.2 Background and Motivation

In this section, some motivating examples are provided of real-life final phase problems.

The US-led military operation "Iraqi Freedom" against Saddam Hussein's regime started in March 2003 and succeeded its initial operational objectives by May 2003. However, it is rarely the case that such operations last for only a few months. Further stabilisation objectives required that the forces had to operate for longer and then to gradually withdraw, and this is what happened until April 2009 (BBC, 2016). From the perspective of the systems' support function of the operations, namely the repair activities and provision of spares for the systems deployed, there were three general phases: the initial build-up, the infinite-time horizon during the stabilisation and the final closing down (Committee on Force Multiplying Technologies for Logistics Support to Military Operations and Board on Army Science and Technology, 2014, figs2–1).

A similar situation took place for the NATO operations in Afghanistan. In October 2014 the US and the UK announced the end of their combat operations, while in December 2014 NATO formally ended its combat missions. Again, the three general phases of initial build-up, infinite-time horizon and final closing down were

present. However, reports highlighted that this final closing down phase was also the “bloodiest” period of its duration (BBC, 2018; NATO, 2017). This finding contradicts the assumption of operational planning that during the closing down phase – the phase that follows the perceived attainment of the planned end-state - the operational demand reduces (see e.g. US DoD (2017b, p. IV-20), (2017a, p. I-8)). On the contrary, in the specific operations in Afghanistan the operational demand did not reduce (BBC, 2018; NATO, 2017). The effects of such a discrepancy on the logistic and repair support of the operations can be considerable. During the final phase, along with other planned changes, a number of the resources including supported systems are withdrawn. Consequently, given the uncertainty about the intensity of operational demand, the support providers need to estimate/forecast the effects of all the anticipated changes on the failure rates. Such a forecast would facilitate decisions on the amount of logistic and repair resources to maintain until the end. Consequently, models used to provide such forecasts need to be able to effectively use any relevant information that is available about the anticipated changes.

A core similarity in the above cases concerned with access to the information needed in the forecast models examined here (Sections 1.4, 1.4, 4.2, 4.5, and 4.6), is the set of relationships among the agents that support the operations throughout their life-cycle. These relationships are usually closer in contrast to those at “arms-length” in which whenever the customer needs to replenish the inventory he/she calls the supplier and places an order (Christopher and Lee, 2004; Christopher and Peck, 2004). Such closer relationships allow the agents to have access to wider relevant data and information (Christopher, 2016, p.156) needed in the forecasting models.

Furthermore, it is not only military support operations, but also today’s businesses that tend to develop closer relationships. The closer cooperation is a general tendency in the evolution of businesses that want to benefit from (global) Support Chains. As Christopher (2016, p. 156) points out “*Today’s business is increasingly ‘boundaryless’, … the separation between vendors, distributors, customers and the firm is gradually lessening. This is the idea of extended*

enterprise, which is transforming our thinking on how organisations compete and how value chains might be reformulated".

Christopher's observation provides a key assumption underpinning this research. For this thesis' purposes, it is assumed that in the cases examined, Support Chain relationships are in place similar to the ones built under availability and performance-based contracts. Such contracts provide the support function as an integrated, performance package set to optimise and meet performance goals, such as the operational availability of a fleet of systems (Mirzahosseinian and Piplani, 2011; D. Nowicki, Kumar, Steudel, and Verma, 2008; D. R. Nowicki, Randall, and Ramirez-Marquez, 2012). Under such contracts, maintenance and servicing of a system is not paid according to the spares used, or the number of workhours, but on the agreed measures' outputs (Mirzahosseinian and Piplani, 2011, p.260). Inevitably, the service provider has access to the customer's data as well as her own and of others that directly participate in the provision of the contract.

Therefore, the move towards closer relationships among the agents that support the operations can give mutual access to the acquisition of the required data and information. As Christopher (2016, p. 156) sates "*Even more importantly it is information shared between partners in the Supply Chain that makes possible the responsive flow of product from one end of the pipeline to another*".

For the uses of this research, "data" is defined as "*facts and figures which relay something specific, but which are not organized in any way and which provide no further information regarding patterns, context, etc.*" (Frost, 2018). On the other hand, "information" is defined as the data which have been "*contextualized, categorized, calculated and condensed"* (Davenport and Prusak, 2000). In essence, information follows data collection and results from their interpretation so that it can be used to support modelling and decision making. Finally, here "knowledge" is defined as a mix of experience, contextual information, and expert insight. It eventually provides a framework for explaining and evaluating new experiences and information (Davenport and Prusak, 2000).

The assumption of closer relationships between the members of a Support Chain is important for the types of models that are dealt with in the present research. Models like regression and Bayesian Networks can benefit both in their development and in their verification and validation from the use of the knowledge of such relationships that can be incorporated from Subject Matter Experts (SMEs) (Field, Miles and Field, 2012; Gelman and Hill, 2007; Heckerman, Geiger and Chickering, 1995). However, such experiential knowledge is not easily acquired from an “arms-length” type of business. On the other hand, if the term learning is used as “the acquisition of knowledge or skills through study, experience, or being taught” (Oxford University online dictionary, 2018a), then, by definition, more opportunities exist to acquire the relative knowledge when the relationships among the participants in the Support Chain are closer such as in the situations referred to earlier.

The present thesis is concerned with the support of systems which are composed of repairable and discardable components and that are deployed to perform operations. At this point, it is useful to also define how the term “Support Chain” (SC) is used here. The term “Support Chain” (SC) is defined as the networked system that has as its Main Useful Function (MUF) (Cameron, 2010) to make systems available for operations, and so, it is composed of the Supply Chain and also of the repair and maintenance activities.

Nevertheless, FPPs within close supplier-customer relationships in an SC are met not only during large military operational deployments such as the ones in Iraq and Afghanistan that were described earlier. In 2009 BAE Systems announced the closure of its Nimrod aircraft production and support plant in Woodford Manchester by 2012. This was due to the UK MoD’s decision to withdraw this old but very capable Maritime Patrol Aircraft (MPA) from its operational status. The final shut-down of the plant actually took place in 2010, two years earlier than initially planned (BBC, 2010; FlightGlobal, 2006), when BAE received a formal “contract termination” notice from the MoD (Kirkup, 2010). However, the anti-submarine and Intelligence, Surveillance, Target Acquisition, and Reconnaissance (ISTAR) missions usually undertaken by the Nimrods were to be only partly covered by other assets (Defence Committee, 2012). For BAE

as the support function provider, the dilemma in this case was similar to the one discussed earlier in the final phase of the large military operations. During the period before the official announcement of the contract termination, as though indications of this outcome were present, decisions on the level of support were required to be based among other information and decision criteria, on forecasts of the failures of the systems and the resulting demand for spares too.

However, the repair and supply requirements during the final period is often more uncertain as compared to the period before that. Using the Nimrods' case as an example, there was a point in time when there were indications on the closure of the operations. From that point, the number of supported systems / planes and their repair facilities that would be kept until the very end were uncertain too. Moreover, given the uniqueness of these Maritime Patrol Aircrafts' (MPA) capabilities in anti-submarine and ISTAR missions (BBC, 2011), the operational requirements were not expected to be changed. That situation included many challenging uncertainties and it was similar to the final phase of the military operations in Afghanistan where the reduction of the support facilities was not followed by a similar reduction in the operational requirements.

Political and economic changes can be the cause of such uncertainties and dilemmas. In 2010 Lockheed Martin announced the closure of its plants in Goodyear Arizona, Akron Ohio, Newtown Pennsylvania and Horizon City in Texas by 2015 due to US government budget reductions. Those plants were producing and supporting the Patriot missile defence systems and the F-35 and F-16 fighter planes (The Seattle Times, 2013).

Even in the commercial world where maintenance support contracts exist, decisions to withdraw the systems supported are not rare (Meridiana, 2018). In 2017 Allegiant decided to replace its fleet of MD-80 commercial aircraft fleet with an A319 and A320 fleet due to the age of the former and the higher reliability and efficiency of the latter (FlightGlobal, 2017). On the same type of problem, Boeing has published studies about commercial aircraft' economic life in which they demonstrate the need for fleet renewals (Jiang, 2013). The resulting problem in the final phase is expected to be similar to the one described earlier. The

companies that have been contracted to participate in the Support Chain of the fleet that is to withdraw, will need to forecast the demand for support until the very last supported aircraft has been removed from operations. Furthermore, due to the increased uncertainties they will need to base such forecasts on models that are able to effectively use data and information from both the operations and their support.

The cases above are examples of the FPP. The FPP describes the context of the current thesis' Research Problem. The FPP creates a decision uncertainty on the support requirements in spares of systems with the following two assumptions. First assumption is that these systems are composed of repairable and discardable components. Secondly, this uncertainty results from the announcement of a finite period during which the systems will be withdrawn from operations. In the FPP context, the decision maker has a single period ahead for which to place an optimum order to fill the inventory. In essence, the ratio of any resupply Lead Time (LT) over the duration of the Final Period (FP) is considered greater than 1. Furthermore, the FPP is differentiated from other similar type of problems (see Sections 2.2 and 2.3) by the expected and assumed increased cooperation and subsequent data and information exchange existing in Support Chains under modern availability and performance based contracts.

The related Research Question dealt with here is to explore what the benefits and difficulties are of developing a number of forecasting models – different versions of Bayesian Networks in particular - that can exploit the SC-wide information and data. These specific models were chosen to be explored due to a number of useful attributes (Sections 1.3, **Error! Reference source not found.**, 2.5), but mainly due to the insights that their graphical structure (Section 4.3) can provide. Their structure represents the joint probability distribution of the modelled factors, and in this way it provides a visual representation of their association. Such a visual output among other benefits, can also facilitate understanding of the interactions existing in a large, complicated system like the Support Chain.

The Bayesian Networks are also compared to two other commonly applied forecasting approaches, the SME adjusted Single Exponential Smoothing (Section 4.6) and Logistic Regression (Section 4.4). The reasons for choosing these two modelling approaches are discussed in the following Section 1.3.

The exploitation of the data using Bayesian Networks aims to facilitate decision making on the level of spares to order and maintain during that final phase period by providing forecasts of the expected demands for spares. As will be subsequently demonstrated, the FPP can have severe adverse consequences in military and commercial operations. Yet despite the magnitude and frequency of the problem, to the best of the author's knowledge it remains a largely under-researched topic.

1.3 Research Design and Methods

This thesis focuses on a special class of models, Bayesian Networks (BNs).

The first reason for choosing to explore the applicability of BNs in modelling the demand in the FPPs, is that, there have been a number of studies investigating their use in related fields including reliability (Langseth and Portinale, 2007), maintenance (Jouffe, Weber and Munteanu, 2004; Weber and Jouffe, 2006), system testing in manufacturing (McNaught and Chan, 2011) and supplier selection (Hosseini and Barker, 2016). However, no studies were found of any application to the kind of logistical support problems outlined here (FPPs). In other words, the current research has not found any study in which the Support Chain that had been formed in order to provide the availability of certain systems deployed during operations is scheduled to be closed, and the decision maker requires an informed forecast of the expected demand for spares during the closing down / final phase period. Furthermore, in the present research, the BNs were developed using the kind of data captured in the logbooks associated with the Support Chain nodes.

From a modelling perspective, the BNs are graphical models which means that a graph or network maps the associative relationships among the variables of the study. The graphical description and its exploratory power is considered a very useful contributor to the present study. In traditional forecast models,

understanding is usually provided through an evaluative explanation of how inputs lead to outputs. So, for example a regression model provides explanatory information to the decision maker via the coefficients of the explanatory variables and their standard errors. On the other hand, when a BN is learned from data, the resulting graph can also reveal influential associations among the variables that formulate the demand context, and which cannot easily be identified by either the experts or the traditional demand forecast models. This benefit is demonstrated in Chapter 7 (Sections 7.2.3 and 7.3.3) and further examined in Appendix B.

Furthermore, a BNs' structure as a joint probability distribution provides the ability to use it not only as a modelling instrument of the relationships among its explanatory variable and the related response, but questions (queries) can be expanded to the relationships among any other subsets of the participating variables (see e.g. Boutselis and McNaught (2018), Nagarajan, Scutari and Lebre (2013), Neapolitan (2004), Pearl (1988)).

Additionally, given their graphical representation, a modeller can conveniently include and verify information that can be acquired from subject matter expertise as well.

In order to be able to provide with a forecast during the final phase period, the decision maker can rely on what is known in the past regarding the way that the demand for spares is related to other variables, like the systems' usage rate, or the environment, and also on what can be known or planned for the following up duration of the final phase. Consequently, in order to compare the forecasts provided by the BNs, a choice was made to additionally develop a logistic regression (Section 4.5) and an experts-adjusted Single Exponential Smoothing (SES) model (Section 4.6). The logistic regression forecast is developed by the use of the variables included in the changing demand context that one would expect to be relevant during the final phase period.

The experts-adjusted model reflects the common industry practice to adjust the forecasts provided by a model – a SES in this case - in order to reflect the decision makers' consideration of the contextual factors' effect on the demand (Fildes,

Goodwin, Lawrence, and Nikolopoulos, 2009; Franses and Legerstee, 2010; Klassen and Flores, 2001).

Figure 1-1 presents an overview of the research design. As discussed further in Chapter 6, a simulation was applied to generate data. These data were used for the development of the forecast models and also for the evaluation of the models' outputs.

The Bayesian Networks (BNs) (for the definition of the BNs' Directed Acyclic Graph (DAG) and Node Probability Tables (NPTs), see Section 4.3) along with the Logistic Regression and the SME adjusted forecasts were developed using the first set of data, the models' development data. This dataset was generated by a single replication of the simulation since it represented the single instance of available data from the support chain that could potentially be collected in real life.

The second dataset - the evaluation data - was produced in order to evaluate the individual models' forecasts. This evaluation dataset represented the FPP period, and in order to evaluate the forecasts across a range of different possible FPP situations, a number of different datasets were produced (Sections 7.2, 7.3). Furthermore, for each different final phase scenario considered, the simulation was run for 100 replications. This was to provide a fairer basis for the evaluation and comparison. Chance variation alone could easily distort comparisons based on a single replication.

A study was also performed to identify appropriate accuracy and accuracy implication metrics to be used for the evaluation of the models' FPP period forecasts. Finally, the evaluation results were compared using these accuracy implication metrics.

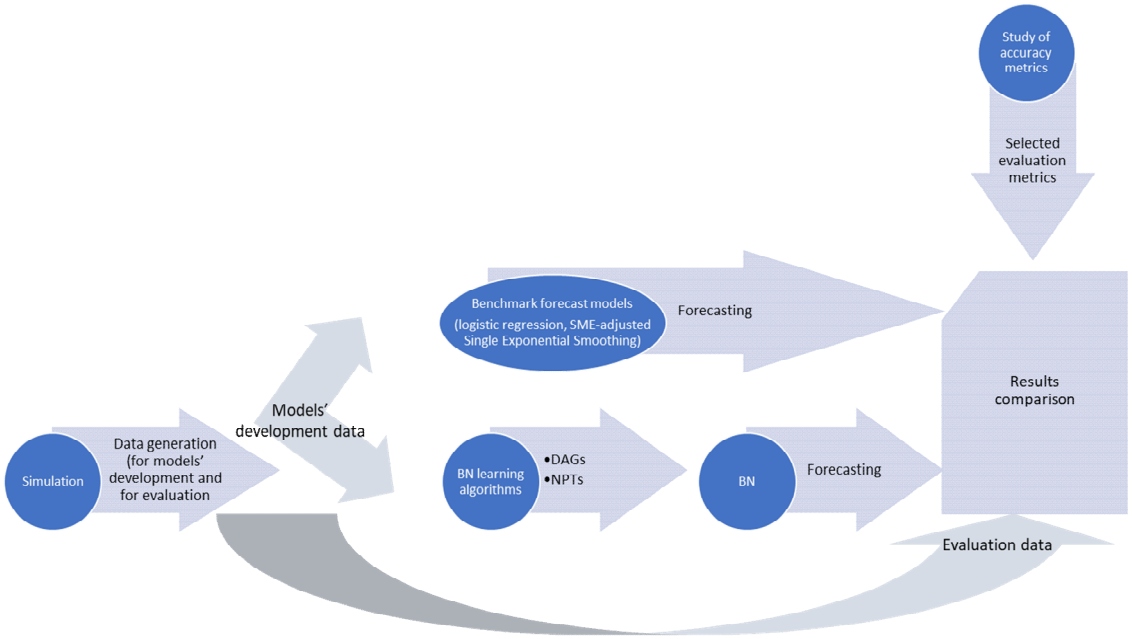


Figure 1-1: Overview of the research design

1.4 Aims and Objectives

In summary, the aim of the present research is to study the demand context which exists during the final phase of a support operation (FPP) and, by the use of Bayesian Networks (BNs), to exploit the data that is available from the different connected nodes of a Support Chain in order to improve spares' demand forecasting for the FPP.

In order to explore the usefulness of BNs in the FPP's context, four different methods of BN structure development were employed and their forecasts compared:

- Bayesian Networks were developed through:
  - Unsupervised machine learning using data collected from the logbooks of the functional and operational nodes of the Support Chain
  - Causal elicitation of the BN structure from experts' knowledge
  - Hybrid development of the BN structure using the expert knowledge as a prior structure and adding a machine learning algorithm that builds upon the elicited structure

- Hybrid development of the BN structure using the expert knowledge and adding a machine learning algorithm that uses that structure as a starting directed acyclic graph (DAG) on which connections are then added, removed and adjusted by the algorithm to increase the likelihood of its structure

For the benchmarking of the BN results, the following forecasting models were also developed:

- A logistic regression for the modelling of the probability of component's failure
- A Single Exponential Smoothing (SES) algorithm that provides predictions to decision makers based on past demand in order for them to adjust given their knowledge of changing demand context factors

In order to make meaningful comparisons, typical performance measures relating to forecast accuracy were reviewed and suitability assessed for this type of problems (FPPs). These included both accuracy and accuracy implication measures (Section 5.2). As it is also shown regarding the implications of the forecasts' accuracy to the effectiveness and efficiency of the spares' inventory, the idiosyncrasy of the FPPs calls for the introduction in the evaluation of some additional measures to the ones commonly applied in the literature (Sections 5.2.5, 5.2.6).

1.5 Thesis Layout

Driven by the research design steps as presented in Figure 1-1, the thesis has been formulated as follows:

In Chapter 2 the literature review provides an analysis of problems similar to the FPP. This review identifies the models that have been applied and the factors that have been taken into consideration in problems similar to the FPPs. Furthermore, the literature is also reviewed in order to gain a greater understanding of the factors that formulate the demand context and that can be used later as explanatory variables in the demand forecast models. To help verify the factors identified in Chapter 2 and potentially identify some additional ones,

primary data were collected via interviews with two relevant experts who were chosen based on their operational background in managing the Support Functions of large military operations of the UK Army and the RAF. More details of these interviews are provided in Chapter 3. This chapter also presents some conceptual models to help an analyst decide which factors might be most relevant in a particular support setting. Chapter 4 presents the Methods applied in the research. The chapter firstly presents the BN models that were evaluated and the respective development methods are presented. The chapter continues by discussing the discretisation of the continuous data that was required in order to be able to use the BNs' structure learning algorithms. Further on, the chapter presents two other modelling approaches that were used as benchmarks: the logistic regression and the Subject Matter Expert (SME) judgmental adjustment of forecasts. Chapter 5 proceeds with the discussion on the performance measures that were applied in order to evaluate the models forecast outputs. In this chapter, suggestions are also made on the accuracy measures' quality and also on additional accuracy implication measures that are needed for the FPP cases. Chapter 6 discusses the stochastic simulation model that was developed to generate the data for the development and the evaluation of the forecast models. One of the conceptual models presented in Chapter 3 is used to help identify relevant factors to include in the simulation model. Chapter 7 describes the two simulated scenarios that were used in order to develop and evaluate the forecasting models. Furthermore, the accuracy and accuracy implication measures of the forecast models for each of the scenarios are presented and discussed. Chapter 8 presents the conclusions from the research, lists the limitations of the study and suggests areas for future research.

Finally, Appendix A includes details of the forecasting models developed, Appendix B discusses a number of observations that were made from the simulated scenarios, while Appendix C includes the preprint of a research paper published as a result of the thesis (Boutselis, P. and McNaught, K. (2018) 'Using Bayesian Networks to Forecast Spares Demand from Equipment Failures in a Changing Service Logistics Context', *International Journal of Production Economics*).

1.6 Conclusions

This chapter has introduced the Final Phase Problem (FPP) as a particular problem in logistics management that has not been adequately addressed in the literature, and which will be the focus of this thesis. In particular, the forecast of the demand for spares during the final phase period will be studied by the use of a number of different Bayesian Networks (BNs). The commonly employed logistic regression model and the SME-adjusted Single Exponential Smoothing. (SES) model will be used to provide baseline comparisons.

The chapter also presented the research design, in which a simulation of the support chain is used to produce data for the forecast models' development, as well as separate datasets for their evaluation. Regarding the evaluation, the research design also includes a study of the forecast models' accuracy and accuracy implication metrics.

2 LITERATURE REVIEW

2.1 Introduction

Forecasting the demand associated with support for the final phase of operations in order to facilitate decision making for this phase is an important problem. However, only some specific variants seem to have been studied in the academic literature. These involve single-period forecasting and include Newsvendor problems and also “Last time buy” / “End of life” problems. The main characteristic that is common to this research and the Newsvendor and the “Last time buy” / “End of life” types of problems is to be found in the specific challenge that the decision maker faces. In all three problems, there is a single period ahead for which the decision maker has to place an optimum inventory order. If the inventory level is lower than the experienced demand then there are underage/shortage costs, while if the level is higher then there are overage/holding costs. Consequently, and as discussed in the definitions of these problems at Sections 2.2 and 2.3, respectively, in all three of them the accuracy of the demand forecast is an important factor that contributes to the optimisation of the decision on how many spares to order at the beginning of finite-time horizon period.

The literature review involved a two-step process. The first step identified and reviewed the modelling approaches that have been used in similar kinds of problems, namely the Newsvendor and the “Last time buy” / ”End of life” problems. The second step then updated the findings of the first step in order to produce a list of factors that have been taken into consideration for the modelling of the demand for spares in the literature. This set of factors is referred to as the demand formulating context factors. Its production is a key part of this chapter since such factors will drive development of the demand models later in the thesis.

Improved understanding of these factors and how they might interact was acquired by interviewing related Subject Matter Experts, as described in Chapter 3.

Figure 2-1 provides a graphic summary of the literature review process along with the associated objectives.

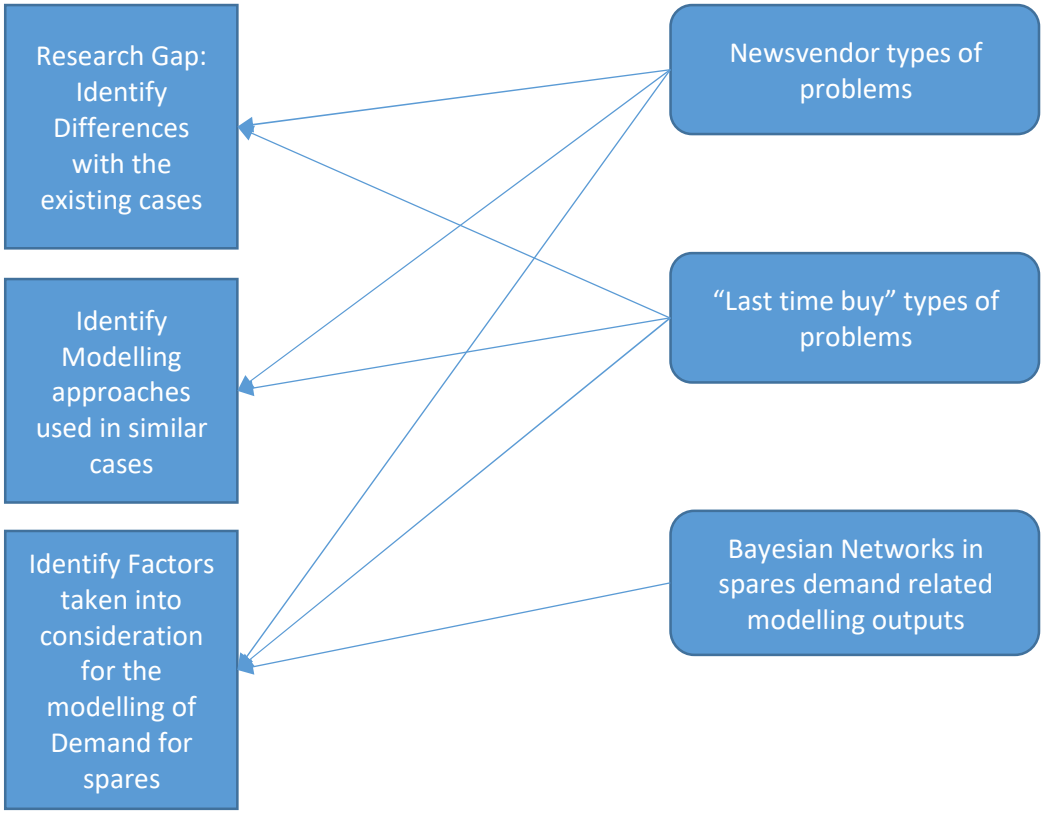


Figure 2-1: The literature review process and objectives

2.2 Comparing the Final Phase Problem to the Newsvendor Problem

The newsvendor problem (NVP) (also commonly known as the Newsboy problem) is one of the classical problems in operations management and has been extensively studied since the pioneering effort of Edgeworth (1888). A recent review of the area is provided by Qin, Wang, Vakharia, Chen, and Seref (2011), while as Alwan, Xu, Yao, and Yue (2016) state, the research on demand forecasting in the NVPs is a topic not well covered in the literature.

The NVP types of problems can be divided into the following two categories: single period and multi-period. In the first category, the NVP is a one-off problem. The decision maker is facing a single situation for which there is no recent background and it will not repeat itself in a following period. Characteristic

examples are support with supplies for a disaster relief operation, a military “fly-away kit” (see e.g. Lancaster (2005)), or deciding how much inventory to hold for the selling period of a new, tailor-made product (Eppen and Iyer, 1997; Lariviere and Porteus, 1999; Zheng, Wu, and Shu, 2016). In such cases, the problem of underage and overage costs is still present, there is not much background knowledge or recent data and the situation is not expected to be repeated, at least not in the sense that the multi-period NVPs are. Consequently, the demand forecast is a challenging task, but the main thing to point out is the fact that in such cases there is no learning that can be extrapolated from the recent past. The decision makers have to rely on their understanding of the single-period’s similarities/analogies to other situations.

On the other hand, in the multi-period category there is a repetition of the same NVP dilemma in a “myopic” way. At each one of the many periods, and despite the fact that there is a continuum among them, each problem is dealt with as a single NVP at a time. Examples of the multi-period category are the newsagents’ decisions concerning the amount of perishable newspapers and magazines to order at the beginning of each period, the grocers’ decisions on the amount of fresh fruit, vegetables and milk to order, or the fashion buyers’ decisions on quantities. In that sense there are directly relevant past data that can be used to help the decision maker make a more contextually informed demand forecast. Furthermore, the decision maker will probably have been accumulating experiences from the previous periods. Consequently, these experiences can result into knowledge on factors like the demand patterns and probably demand’s relation/association to certain factors (Rekik, Glock and Syntetos, 2017). The existence of useful knowledge coming from the demand-patterns’ repetitions can also be inferred from the discussions of Alwan, Xu, Yao, and Yue (2016) and Özer, Uncu, and Wei (2007) who observe that NVP demand is rarely Independent and Identically distributed (IID) and that there is autocorrelation between the periods.

The Final Phase Problem (FPP) cases cannot be placed clearly under either of the previously mentioned categories. The closing down of an operation is a one-off challenge and it will not repeat itself in the multi-period sense, so it has some

of the demand-prediction challenges of the single-period case. On the other hand, there are recent background data from the build-up and the infinite-time horizon phases. Consequently, a level of learning has accumulated as in the multi-period cases, but given the possible changes that will follow during the final phase it is not certain that the demand will keep the same pattern nor that the previously induced associations will still be the same..

The earlier mentioned observation that the FPP cases cannot be clearly categorised as either single period or multi-period NVPs, has an effect on the data that can be used to produce a demand forecast and thus on the spectrum of demand models that can be applied. In the single period, a way followed to collect data is if some kind of postponement for the time of the decision making can be accommodated until some new demand data are available in order to inform a Bayesian update (Eppen and Iyer, 1997; Hill, 1997; Lariviere and Porteus, 1999). Zheng, Wu, and Shu (2016) studied the potential postponement of an order in different supply scenarios and concluded that a postponement has the negative effects of increased costs of purchasing and of shorter ordering times. Their conclusions can be related to the cases examined in the present research. During the final phase of an operation, there could probably be an opportunity to place an order after the beginning of the period of interest at a premium and thus wait to collect valuable data. However, this can be a choice not taken mainly due to the very long lead times, an assumption that has often been made in a number of multi-period NVPs as well for each individual period (Rekik, Glock and Syntetos, 2017). For the single period NVPs where no update is chosen then the prior estimate for the demand distribution is provided by experts' judgement using engineering thinking and/or their experience from other similar cases is commonly suggested alternative to provide a demand forecast (Berk, Gürler, and Levine, 2007; Ding and Gao, 2014).

On the other hand, in the multi-period NVPs there is recent evidence of similar situations due to the repetition of the similar periods. In such cases, there is no need to wait for some data to be collected during the period under consideration as it was the case in the single period NVPs. The past periods of the multi-period NVP help in that both data and learning exist, and thus demand forecasting can

be informed from these periods. Nevertheless, Bayesian updates are not uncommon in the multi-period NVP cases either. Berk et al. (2007) studied the use of conjugate priors in developing expressions for the exact posteriors for a number of demand distributions. Choi, Li, and Yan (2004) present for the multi-period NVPs the same observations as Zheng et al. (2016) do for the single period NVPs, i.e., that the Bayesian demand updates can result in a better forecast and a lower demand uncertainty but at the expense of higher costs (and less lead time).

Another demand forecasting approach that has been used in the multi-period NVPs is the time series statistical models (Alwan et al., 2016; Carrizosa, Olivares-Nadal, and Ramirez-Cobo, 2016). However, these cannot cope well on their own if there are substantial changes expected to take place in the forecasting horizon (Dekker et al., 2013; T. Y. Kim, Dekker, and Heij, 2017), and this is how the situation can be in a number of cases during the final phase of support operations. In order to be able to accommodate the anticipated changes within the demand forecasting, decision makers of NVP problems have occasionally applied regression/econometric models (for an example see Polatoğlu (1991), and for an overview of the respecitve literature see Qin et al. (2011)). An approach that is often used in order to accommodate changes is to include SME judgemental adjustments to either the statistical demand forecasts or the order level suggested by the NVP optimisation models (Rekik, Glock and Syntetos, 2017).

Timing categorisation of the NVPs (as a multi-period or as a single period) as compared to the FPP cases that are examined in the present research is not the only difference that can affect the data and information which can be used in order to develop a demand forecast model. Another important difference is in the types of items considered. The inventory of interest in the NVPs is usually about commodities whose demand does not depend on one another apart from the multi-item/multi-product NVP cases where different commodities either share the same shelf in the inventory (Martín-Herrán, Taboubi, and Zaccour, 2006; Urban, 2002), share the same budget or capacity in general (Abdel-Malek and Montanari, 2005; L. H. Chen and Chen, 2010; Khouja and Mehrez, 1996; Luo, Wang, and Chen, 2015; Vairaktarakis, 2000), are cross-selling/substitutes

(Casimir, 2002; Huang, Zhou, and Zhao, 2011; Kraiselburd, Narayanan, and Raman, 2009; Zhang, Zhang, Zhou, Saigal, and Wang, 2014) or they are a single type of items that compete for a market share (F. Y. Chen, Yan, and Yao, 2004; Huang et al., 2011; Zhao and Zhao, 2016).

The FPP cases examine repairable and discardable parts that do not only compete on the shelf or for a budget, but they also interact inside the systems in which they are installed as components and also through the number of systems that operate and are supported (Kennedy, Wayne Patterson, and Fredendall, 2002). The dependencies created due to the interactions of the components inside each system and among the systems, present a fundamental idiosyncrasy of the present research. These interactions affect the factors that need to be considered for the forecasting algorithms. For example, an unavailable part from the inventory affects the usage rate of the whole system and thus of other components as well (Behfard, Van Der Heijden, Al Hanbali, and Zijm, 2015). The identification of the dependencies among the parts is also one of the core findings of the present research and it demonstrates the benefits of studying the forecasting problem through a BN even if one eventually decides to build another model and/or rely on experts' adjustments (Appendix B).

A special NVP case in which spares for systems are taken into consideration is the "Fly-away Kit" (Sherbrooke, 2004 pp 214-215) or similarly the "Endurance" scenario (Systecon, 2015). However, in each of these applications a Poisson-family distribution is assumed with no discussion on how its mean has been acquired. Additionally, both of these cases are single periods without any other periods considered prior to them.

An additional difference of the NVPs as compared to the FPPs and which is directly related to the types of items and the related data examined, is about the industry sectors and types of business that each problem considers. Retail industries are the main business sectors for which NVP's studies have shown interest, while the present research is more relevant to the systems manufacturing and service support industry. What distinguishes the latter from the former is that systems support business are interested in the systems

themselves as well as in the provision of supplies in spares ranging from simply producing and supplying the spares whenever a customer places such an order, to delivering availability or performance based contracts (Behfard et al., 2015; Mirzahosseinian and Piplani, 2011; D. Nowicki et al., 2008; D. R. Nowicki et al., 2012; Ruud H. Teunter and Fortuin, 1999). On the other hand, retail businesses are interested only in delivering the customers' ordered commodities, which is a challenge different to the spares' demand forecasting since in retail business the modeller needs to forecast the buying behaviour of the customer.

The previous two examined NVP idiosyncrasies, i.e. the type of items being commodities and the type of industries being the retail business, has had a driving effect to the direction that has been followed in the literature regarding the identification and use of demand factors / predictors in regression and econometric models. Particularly, in retail, the examined influential factors have been the selling price, the supplier's price discount, the advertisement size etc. (Boutselis and Mcnaught, 2014; Khouja, 1999; Qin et al., 2011).

On the other hand, in the FPP cases the production and service support of systems and their demand for spare parts does not depend only on their price and attractiveness to the customer. Apart from the interdependence among the components that are engineered on the system, there is a dependence on the attributes of the installed base – that is the number of systems that have been installed and need to be supported (Dekker, Pinçe, Zuidwijk, and Jalil, 2013; Jalil, Zuidwijk, Fleischmann, and van Nunen, 2011; Jin and Liao, 2009) - as well as on qualitative factors like the environment in which the systems are operated and in which the support is provided, the levels and the quality of maintenance etc.(Boutselis and McNaught, 2018; Sherbrooke, 2000, 2004, pp.291–299). This plurality of possible interactions makes the number of candidate influential factors that formulate the demand context quite large and the problem of producing a demand forecast by either an expert's judgment, or a type of model that uses these factors as predictors (e.g. regression, Neural Network, Random Forests etc.), or a combination of the two, quite challenging.

FPPs require access to data and information on multiple influential factors related to the demand. Such a requirement is not a problem that is identified in the literature for the NVPs, and especially the multi-period NVPs. In the NVPs, the demand forecast utilises the demand data patterns from the past periods. This fact makes the efficiency of use of time series statistical models an advantage of the NVPs as compared to the FPPs. On the other hand, for operational systems and their component parts and the industry that is related to their production and service, there is a need to have access to data such as those related to the installed base (Dekker et al., 2013; Pince and Dekker, 2011; Van Wingerden, Basten, Dekker, and Rustenburg, 2014) and the usage rate of the systems. Inevitably, this requirement calls for access to data owned by multiple agents, those cooperating at different levels in the Support Chain of the deployed systems. This is not always achievable especially for “arms-length” contracts that have low level of cooperation. On the other hand, in availability or performance based contracts (Mirzahosseinian and Piplani, 2011; D. Nowicki, Kumar, Steudel, and Verma, 2008; D. R. Nowicki, Randall, and Ramirez-Marquez, 2012) the required access to data is easier since the resulting relations tend to evolve towards extended information sharing (Christopher, 2016, p.156; Christopher and Lee, 2004; Christopher and Peck, 2004).

In an availability contract, in order to be able to provide the required level of spares, the contract owner needs to consider the support and supply capacity since they both affect the availability of the operational systems and thus their usage rate and consequently the demand for spares (Lau and Song, 2008; Sherbrooke, 1967, 2004). Moreover, support through repair is important since repairable items tend to comprise the largest part of complex systems’ inventory value (Van Kooten and Tan, 2009; Sherbrooke, 2004, p.6; Syntetos et al., 2009) while related industries tend to allocate an increasing value on service part inventory investment (M. A. Cohen, Kleindorfer, Lee, and Pyke, 1992; M. a. Cohen, Zheng, and Agrawal, 1997; Johnston, Boylan, and Shale, 2003). Under the assumption that the decision maker will have access to different areas to collect data, the present research has focused on the exploitation of the records kept in logbooks of the different nodes of the Support Chain and used that type

of data to develop different demand forecast models. This assumption also provides the convenience of having direct access to demand data and thus reduces the need to approximate them with sales data as is usually the case in demand forecasting studies (Syntetos et al., 2009; Syntetos, Nikolopoulos and Boylan, 2010).

A notable study of the NVP (multi-period) demand forecasts is that of Rekik et al. (2017). They present a study of judgementally adjusting either the demand forecast or the order to be placed after a “signal” has been received by the decision makers, while the first two moments of the demand’s distribution are estimated from historical data. In their work the decision maker does not have to wait for new demand data. However, there is a “signal” factor which the decision maker believes is associated with the demand during the single period under consideration and also that this “signal’s” probability of being correct can be different from 1, i.e. the signal is imperfect. An example that they give is of the observed high sales of a mobile phone being associated to possible sales of a tablet of the same brand. However, as they state, if the customers are not eventually satisfied by that specific product, then its high sales signal might be false and misleading. Furthermore, using the opportunities offered from the rolling periods of the multi-period problem they suggest that learning can increase if records are kept on the price and cost of the product, which are factors that affect the demand. Additionally, Rekik et al. (2017) suggest that the usual challenge of acquiring access to data could be overcome if data owners would cooperate by raising the awareness on the advantages of improved forecasting.

2.3 Comparing the Final Phase to the “Last time buy” Problem

Another area of research that is related to the FPPs is the one in which the problems examined have names such as “Past-model” or “All-time requirement” of replacement parts (Fortuin, 1980, 1981; Moore, 1971; Ritchie and Wilcox, 1977), “End-of-life” (EOL) service (Kleber, Schulz and Voigt, 2012; Leifker, Jones and Lowe, 2012, 2014; Pourakbar, van Der Laan and Dekker, 2014; Teunter and Fortuin, 1998, 1999), “Final order” (Van Kooten and Tan, 2009; Teunter and Haneveld, 1998, 2002), “Discontinued product” (Hong, Koo, Lee, and Ahn, 2008),

“Last time buy” (LTB) of spare parts (Behfard et al., 2015; Krikke and Van Der Laan, 2011), “Post product life cycle” (Inderfurth and Mukherjee, 2008) and “Spare part procurement after End-of-production” (Inderfurth and Kleber, 2013). These problems are all “single period” decisions about spares’ inventory for repairable systems and repairable parts, and they involve decisions concerning the final phase of the operations that follows the building up and the steady-state phases. B. Kim and Park (2008) highlight the importance of forecasting the demand to the decisions during this final period of interest. Due to their resemblance to the problems of the interest of this research, the literature was investigated in order to see details on the demand forecasting models used and the factors considered.

However, in the literature on these types of problems (LTB/EOL) that the author examined, the “final” decision under consideration was either mostly seen from the perspective of the “seller” (spares’ provider), or in fewer cases from that of the “buyer” (Teunter and Haneveld, 1998). The importance of this distinction is highlighted in Leifker et al. (2012) where they claim that the manufacturer, unless he/she has a close working relationship with the customer, he/she is likely to have access only to the part and product failure rates and maybe also to a probability distribution of the number of products still in operation. This observation stresses the potential importance of having access to additional information that could be acquired from a stronger relationship between the seller and her buyers. A similar point is deduced by (Kennedy, Patterson and Fredendall, 2002) in their spare parts inventory management review. The authors mention that the need for spare parts is dictated mostly by the maintenance policy rather the customer usage. This statement suggests two things. Firstly, that both maintenance policy and customer usage have an influence to the demand for spares, and secondly that since such a knowledge needs access to service and usage related data, in order to increase such a knowledge there is a requirement to have access to both the “customer” and the “service provider”.

Furthermore, the related problems that the author found in the literature review were more in the sense that the decision maker (either the buyer or the seller)

wants to proceed to a spares purchase as an "insurance[1]" provisioning action during the LTB/EOL of the purchased systems (Geurts and Moonen, 1992; Teunter and Haneveld, 2002) which implies an assumption of a consequent reduction in the usage rate of the systems as well (Hong et al., 2008; Moore, 1971; Ritchie and Wilcox, 1977). This assumption stems from the fact that in the examined literature the general context is that a product/system that is introduced to the market has a limited lifetime and then at some point a decision is made to stop the production of the system and of its spares (probably due to an introduction of newer technology) while the sold systems are still operational. The consequence of such a scenario is that the resulting LTB/EOL modelling approaches try to accommodate an unquestioned expectation of a gradual reduction in the demands for service and thus for spares.

Moore (1971) appears to be the first to have tried to model the demand forecast in such a context. Moore claims that the application of either uninformed time series like exponential smoothing or regression models that are based on variables with data accessible to the "seller", i.e. the number of original equipment sold, the earlier warranty claims or service calls and the ratio of part to market value, cannot cope well with the problem of forecasting demand. Moore studied sales data from different types of spare parts from the automotive industry during the final phase of the support provision due to the replacement of the vehicles with a new model. He observed that when these spare parts sales data are plotted on a base-10 logarithmic scale they tend to follow either a parabolic, or an elliptic curve, or be linear. He therefore suggested that if the decision makers collect spare parts sales data as a proxy for spares demand and identify the peak (which indicates the start of the final phase) of the sales then, after a few data points one of the three curves can be fitted and thus (deterministic) sales' forecasts can be projected into the future. However, his statement about the regression on equipment sold refers to their "original" number and not to the number that is operational and needs support during the final phase.

[1] This differs from the concept of "safety stock" in that it is not maintained to absorb fluctuations in the demand, but rather it is kept "as an insurance" to cover for a likely disruption of the supply chain

Furthermore, Moore's observations are applicable if there is time for enough data points to be collected and thus an equally extended final phase.

Moore's observation has been adopted by Fortuin (1980, 1981) who built a range of deterministic curves that can be used to calculate the level of the final order needed to attain a set customer service level. Moreover, Dombi, Jónás and Tóth, (2018) assumed a concave unimodal demand time-series and developed a deterministic Demand Model Function with short-term fluctuations around it. Nevertheless, these cases still address the long-term LTB/EOL problem from the sellers' perspective that needs to take an aggregated view of the total sales and their life-cycle in order to carry out capacity planning.

On the other hand, Ritchie and Wilcox (1977) take into consideration that there might not be much time to collect data before the decision is made and suggest a deterministic calculation of the expected "all-time future demands" for spares during the final period by incorporating as parameters the number of systems that are subject to failure (the systems that are currently operating), the failure rate of each system's component and the probability that the component that fails will not be replaced. Ritchie and Wilcox (1977) suggest that the parameters' values can be estimated from past data through minimising the sum of squared errors approach. An interesting feature of this work is the recognition of the effects of the repair policies on the experienced demand. On the other hand, they assume independence among the components/parts, an assumption that might be true on an aggregated level of multiple customers (Inderfurth and Kleber, 2013). However, at a more detailed level like the one of an availability contract, and as Inderfurth and Mukherjee (2008), Kennedy et al. (2002) also suggest, different parts failures depend on each other mainly due to the limited repair and support resources but also due to the same operational requirements (see also Krikke and Van Der Laan (2011)).

Hong et al. (2008) present an extension of Richie and Wilcox's stochastic model in which they also include the potential decrease in the systems' population by discarding them. They include three factors in their demand forecasting models which they combine to formulate what they call "effective demand". These factors

are the failure rate of the component which they combine with the probability of having this component replaced on the system (modelled as a decreasing function of the time) and with the discarding time of the component. This effective demand is then combined with the number of product sales to give the mean and the variance of the total demand during a considered period of time (typically a year). Nevertheless, given that their model is from the seller's perspective and considers the whole installed base, just like Ritchie and Wilcox (1977), the method of Hong et al. (2008) needs to assume that the systems and parts have IID failure time distributions and therefore that the part failures are independent.

Teunter and Fortuin (1998) provide a case study about demand during the "End of Life" (EOL) of a set of components. The authors provide methods and algorithms for the estimations of the final order of Philips electronics spares for appliances sold to private individuals. They produced demand forecasts that they extracted from historical data kept by the company, using a heuristic that estimated the probability that in the following n years the demand in spares for a specific part will exceed the demand in the past 12 months by a factor k $(k = 1, \dots, 20)$. The 12 months were chosen due to the way that the data were recorded and also due to the authors' assumption that most final orders are made about 12 months after the date of the sale. A point worth mentioning from this research is that the authors suggest that they would expect differences between the demand behaviour of a professional customer (e.g. an organisation) and individuals, thus highlighting the importance of having information about the repair policies followed by the customer and about the way the systems are used. Furthermore, in their demand data study, they identify a number of component characteristics that influence the demand, even though they do not use this information in their demand forecast models. They find that the coefficient of variation of the demand decreases when the expected demand increases, that the demand rate decreases faster for the more expensive products which suggests the potential influence of the customers' repair policies and that the demand rate change depends on the type of the component (if it is "standard" i.e. common in different products, or "specific" to only one product) and if that component is in its building up phase or at the end of its life (EOL). Furthermore,

they state that they were not able to show a significant relationship between the demand rate with how long ago the component has been introduced in the market, the type of the component (mechanical or not) and its specific description (e.g. transformer, tube etc.).

Krikke and Van Der Laan (2011) study the relation of the demand during the final period with a number of factors for which the modeller can have access to data. More specifically, they suggest that the demand for spares changes (i.e. is non-stationary) as the size of the systems supported (the installed base) reduces as well. They assume that the demand follows a Poisson process with a rate of $\mu \times IB(t)$ suggesting that there is a linear relationship of demand with the size of the installed base (IB). However, as shown (Section 7.3 and Appendix B) the dependence is neither linear nor it is always positively correlated to the IB. Apart from the size of the installed base, the demand is also governed by a number of factors including the usage rates, the repair resources and the repair policy and this fact can cause significant non-linearity.

2.4 Factors for the Definition of the Demand Context

A list of the factors that were identified in the literature examined above and that have been considered as contributing to the formulation of the demand context is as follows:

1. The costs of acquiring a spare (and other related costs e.g. holding costs)
2. The customer's usage of the system and its components
3. The repair policies and the ad-hoc repair decisions
4. The duration of the single-period under consideration
5. The number of systems (size of the installed base)
6. The failure rates of the systems' components

Apart from the factors identified in the literature on NVPs and LTB/EOL related research, the literature on spare parts logistics was also explored in order to expand the understanding of which factors can contribute to the formulation of the demand context.

Sherbrooke (2000) presented an explanatory study on the effect of flying-hours duration on the demand level for spares on aircraft. His objective was to explore and then explain why during the Operation "Desert Shield / Desert Storm" in 1991 the predicted levels of demand for spares was about the same as those in peacetime despite the fact that the duration of flights was longer during the Operation. He analysed more than 700,000 sorties and found that the following features showed statistical significance according to his data:

- Sortie number during the day, in case each aircraft flew multiple missions during a single day. This probably is a repair decision related factor in the sense that the repair operations could have been deferred until the end of the day and thus any demand for spares would be recorded against the last sortie of the day
- Mission type
- Location
- Sortie duration

Furthermore, Sherbrooke also intended[2] to perform a controlled experiment that would additionally include the following features:

- Aircraft material condition
- Aircrew proficiency
- Deferred maintenance (subset of which is the Sortie number during the day, since the maintenance is deferred if systems are to be used for other missions within a day)

What can be observed is that Sherbrooke's extensive support experience considered *systems engineering factors* like the material condition, *operational factors* like mission type and sortie duration, *support factors* like deferred maintenance and soft factors like aircrew proficiency. This consideration, suggests that these three categories of factors can also be used to identify them when in search for what to include in a demand forecast model.

[2] It was eventually decided not to proceed, mainly due to the costs associated with such a detailed effort

Furthermore, there are a number of implied relationships in the literature that seem to reveal or verify factors that can contribute to the formulation of the demand context but are not expressed as such. Such interesting relationships can be inferred from the work embedded in Multi Item Multi Echelon (MIME) spares optimisation algorithms (Feeney and Sherbrooke, 1965; Sherbrooke, 1967, 2004). Algorithms like VARI-METRIC (Sherbrooke, 2004, Appendix A) are based upon Palm-Khintchine's theorem. In these algorithms' use, if demand for an item is a Poisson process with annual mean m and the repair time for each failed unit is independently and identically distributed according to any distribution with mean T years, then the steady-state probability distribution for the number of units in repair has a Poisson distribution with mean mT. However, this theorem assumes an infinite number of independent renewals which implies a large amount of support resources and perfect repairs as well. Therefore, the number of support resources and the quality of the repair activities (a qualitative factor) should be considered for inclusion in the demand context. Consequently, the following factors should be considered for addition to the set:

- Number of support resources (mechanics, spare parts, repair equipment), which can be categorised under the *support factors*
- Quality of repair, which can also be categorised under the *support factors*

Moreover, not only Palm-Khintchine's theorem but VARI-METRIC models themselves imply the existence of additional demand context factors. These models balance the inventory held in the Support Chain's (SC) depots and pipeline to the repair and resupply times in order to minimise the expected number of backorders in spares demands. Consequently, the:

- Time to repair
- Time to resupply
- Inventory levels
- Inventory ordering policies

should also be considered for inclusion in the set of demand context factors and can again be placed under the *support* category.

An additional area of literature that was considered relevant was that concerning the application of Bayesian networks to reliability and maintenance modelling. First of all, as was just shown, reliability and maintenance are associated with the demand for spares and thus, the author expected in this way to expand his understanding of the spectrum of factors that have been considered. Secondly, an objective of the present research is to explore the applicability of BNs to the FPPs, and thus it is required to know how they have been applied in similar problem-areas.

2.5 The Use of Bayesian Networks in Spares Demand Modelling

It appears from the literature review that BNs have not been used in the modelling of NVPs or LTB/EOL types of problems. However, as Kennedy et al. (2002) intuitively suggest, the demand for spares depends on the number of failures experienced and on the maintenance policy applied. Furthermore, bibliometric studies (Medina-Oliva et al., 2009; Weber et al., 2012) have shown a high interest in the use of BNs in fields related to reliability and maintenance. Consequently, the author was keen to search the literature and evaluate their applicability to the problems of the present research, and through that to also identify the factors that have been included in research for the modelling of reliability and maintenance.

As Langseth and Portinale (2007) point out, BNs have a number of interesting attributes that make them attractive for their application in reliability and in maintenance. They can model complex systems and include complex dependencies among the modelled variables, while the variables themselves can be multimodal. Furthermore, BNs provide a visual representation of the dependencies among the variables and thus help in gaining insight to the modelled system (Boutselis and McNaught, 2018). Moreover, they can conveniently combine diverse data including historical records and experts' knowledge (Weber et al., 2012). Additionally, BNs have the ability to update calculations according to existing evidence, while this evidence can be certain/hard, or not fully known i.e. "virtual" or "probabilistic" (Mrad et al., 2015; Neil, Fenton, and Forey, 2001). This latter modelling benefit is quite important given the type of information that can be available on the future value of demand-

influential variables (Rekik, Glock and Syntetos, 2017) like the environmental conditions, or the usage rate. On the same topic of not fully known evidence, there can be cases in which the modeller has access to datasets that include variables which are indeed influential (e.g. the number of systems that are supported) and that are included in the training set during the development of demand forecast models, but the values of these variables might or might not be available at the time that the decision maker needs to make the prediction. In such cases, models like Neural Networks cannot provide a forecast if the values of all the variables are not known with certainty, while BNs can do so by marginalising out the variables for which the values are not known. A final attribute that makes BNs attractive is that they can also be used for diagnostics (McNaught and Chan, 2011) and thus can help investigate the demand context, and also for prognostics and thus, through the use of (even imperfect) measurements provide evidence for the systems' condition (McNaught and Zagorecki, 2009).

As was mentioned earlier (Section 2.3), demand for spares is caused by fault incidents and by the related maintenance policies. These two initiating points have been at the centre of the analysis by those researchers that advocate the use of BNs in fields related to reliability and maintenance. However, those researchers have mainly focused on a single system and its exogenous (operating environment, maintenance policies/decisions etc.) and endogenous (engineering structure of the system) factors that are associated with the failures. Consequently, as is discussed in more detail later, the demand context has not explicitly included the interactions of the system with the rest of the installed base, nor the related SC functions that can affect the availability of the systems and through that the realised usage rate. In this research, it is shown using experts' knowledge (Section 3.2), simulation experiments and the demand forecast models (Sections 7.2 and 7.3) that such factors are also highly influential and need to be considered in the demand forecast models.

Langseth (1998) was one of the first to compare standard statistical methods to BNs in the analysis and prediction of mechanical equipment's survival times. He used a dataset from the "Offshore and Onshore REliability DAta" database

(OREDA) (Norwegian Petroleum Safety Authority, 1997). Unlike most of the other researchers after him, Langseth used an unsupervised learning algorithm (Bayesian Knowledge Discovery (BKD) (Ramoni and Sebastiáni, 1997)) to generate the most probable BN structure that can be learnt from the data (Heckerman, Geiger, and Chickering, 1995), while in order to calculate the Node Probability Tables (NPT) without having to discretise the continuous variables, he applied BUGS (Gilks, Thomas, and Spiegelhalter, 1993; Lunn, Jackson, Best, Thomas, and Spiegelhalter, 2013) on the chosen BN structure. His BN model was used to investigate the associations among the variables while he also compared it to a standard modelling approach for the system's survival time forecasts, a Cox-regression. In order to formulate the failure context and model it with the BN, he used twelve different recorded attributes from the OREDA dataset, namely: the specific unit's time to fail (as the response variable of interest), the installation's ID, the geographic location, the system's code, the exposure to the environment, the subunits, the design class, the manufacturer, the operating mode, the planned and the actual preventive maintenance and the severity of the failure (see also Langseth, Haugen, and Sandtorv (1998) for more details on some of the used attributes). Furthermore, as Langseth concluded, variables like the aggregated operational time that describe the historical performance were not considered since each unit/case had been recorded for only a period of time and thus exponential survival times were assumed. His comparison did not show a predictive dominance of the BN over the baseline Cox-regression. However, it highlighted the increased understanding offered by the BNs through the ability to visually observe the formed associations among the variables in the BN structure, something that was also identified and presented in Appendix B of the present research.

Additionally, Langseth highlighted an additional benefit of developing the BN using an unsupervised algorithm and which was also independently observed in the present research work (Appendix B). He saw that two of the variables ("Actual Preventive Maintenance", and "Planned Preventive Maintenance") that were shown by the BN to have a direct connection with his variable of interest ("Time to Fail") were not significant at the10% level in the Cox-regression. This fact,

combined with the SME's knowledge that these variables should have been included in the model, highlighted the benefit of developing a BN using the available data so that a better understanding is gained. Consequently, using the knowledge gained by the BN's structure along with any existing scientific knowledge the final model (whether it is a BN or regression) can be built.

Kang and Golay (1999) used a BN as a diagnosis tool that provided advice for the Operational Availability of complex engineering systems, a nuclear power system in their cases, and compared it to the conventional, rule-based expert systems. Given that the main aim was to have a diagnosis tool, the authors adopted an experts' knowledge elicitation approach in order to develop a BN combined with decision and utility functions - a Bayesian influence diagram. The factors that they included were those provided through sensor readings that they used in order to formulate the probability distributions of the failure modes of a monitored system/component. Given the aim of the model to be a diagnostic tool that uses information inferred from sensors' readings, the application was focused on a single, isolated system, without considering the context in which it operates, like the energy fluctuations or the service policies and the related support resources. In that sense, in an extended problem scope like the ones originating at the contexts of Support Chains, such factors would need to be incorporated in the model development process.

Sigurdsson, Walls, and Quigley (2001) attempted to formalise the development and use of BNs for reliability modelling during the design of the system, i.e. before it is put into production and operation. Inevitably, since the intention was to deal with a not yet operating system the directly available data would be limited, so, in order to build the model, the authors suggested to resort to experts' knowledge of other similar systems. Furthermore, they relied mainly on four high-level, qualitative factors, namely the suitability of the design process, whether the standards of the manufacturing are satisfactory, if the screening and test coverage are suitable and if the working environment is favourable. However, even if these factors are of high importance, when hard data on more detailed influential variables will be available after the production and operation of the system, the BN model will eventually need to be redesigned to incorporate them.

Jouffe, Weber and Munteanu (2004) suggested the consideration of factors like the component's age, the maintenance operations and the changing environmental conditions are required as influential factors that can be included in the model of dynamically changing reliability parameters. The authors advocated the use of dynamic BNs as a means to include the degradation of the system through time as well as the changing environmental contexts.

A detailed study of the factors responsible for the failure rate of a system is provided by Jones, Jenkinson, Yang, and Wang (2010). These authors included the age of the component, its life expectancy, the inspection interval and its success rate, the temperature in which it operates and other environmental factors (e.g. electrical power variation etc.), but also the competence of the inspection, of the maintenance regime and of the operating personnel. The consideration for inclusion of influential factors like the competence of the personnel show the tendency to include in the model such qualitative estimates of human skills, something that was also found in the studies of Sherbrooke (2000), (2004). The importance of such qualitative factors in the formulation of the demand context was also verified from the SME interviews presented in Section 3.2 of the current research. Furthermore, Jones et al. (2010) made another important observation that the present research has also (independently) identified. (Appendix B). They suggested that within the influential factors, the modeller may need to include the failures of seemingly insignificant equipment or components, but which have a knock-on effect on the overall failure rate. What this suggestion implies is that it is important to consider those components with the higher failure rates in the modelling of the failures of the other components, mainly because the availability and service activities of the former have a great effect on the repair delay times of all the rest of the parts.

Doguc and Ramirez-Marquez (2009) offer a different BN reliability modelling approach, making use of the K2 unsupervised structure learning algorithm (Cooper and Herskovits, 1992) with historical data. The authors question assumptions such as the experts' unbiasedness and knowledge completeness of really complex systems, and counter-suggest the use of historical data (given of course that the datasets are sufficiently large) in order to eliminate the need for

the use of experts in the development of a BN. Like Langseth (1998), Doguc and Ramirez-Marquez suggested using the BN to find associations among system components. However, in their considerations they included only the components' failures and not any of the other quantitative or qualitative variables suggested in the previously mentioned studies.

A meta-analysis of the use of BNs for the modelling of systems' reliability and maintenance was provided in Medina-Oliva et al. (2009) . They suggested a set of qualitative and quantitative factors to include in the BN models, such as exogenous maintenance action events, production levels, environmental conditions, technical, organisational, informational, decisional and human factors related to the system and its use, and degradation factors such as service time, age, number of repair requests and the planning and execution of maintenance actions.

In another meta- analysis, Weber et al. (2012) suggested that the integration of factors like the technical, organizational, informational, decisional and human and also their impacts on the system's good functioning is an underdeveloped area of research. They also stress the necessity to utilise several sources of information for developing a BN model.

From the earlier discussion on the uses of BNs, two general observations can be made. Firstly, a number of qualitative and quantitative factors have been considered within the three already identified categories of *engineering*, *operational* and *support factors*, and secondly that one more category has been identified, that of the *environment*.

In more detail, the list of factors that could be considered for inclusion in the demand formulating context set and that have been identified until now, is in the following Table 2-1:

Table 2-1: Factors that contribute to the demand, as identified in the literature

SN	Factor identified	Category	References
1.	The geographic location where the systems are operated	*an environmental factor*	Langseth (1998), Sherbrooke (2000)

SN	Factor identified	Category	References
2.	The exposure to the environment	*an environmental factor*	Sherbrooke (2000), Medina Oliva et al. (2009), Jones, Jenkinson, Yang, and Wang (2010), Jouffe, Weber and Munteanu (2004)
3.	The environmental factors in general (e.g. electrical power variation etc.)	*an environmental factor*	Teunter and Fortuin (1998), Jones, Jenkinson, Yang, and Wang (2010), Jouffe, Weber and Munteanu (2004)
4.	The operating mode, and generally the customer's usage of the system and its components[3]	*an operational factor*	Teunter and Fortuin (1998), Langseth, Haugen, and Sandtorv (1998), Langseth (1998)
5.	The competence of the operating personnel	*an operational factor*	Jones, Jenkinson, Yang, and Wang (2010), Sherbrooke (2000)
6.	The system's code / particular configuration of the system	*a system's engineering factor*	Kennedy, Patterson and Fredendall, (2002), Krikke and Van Der Laan (2011), Langseth (1998), Sherbrooke (2000), Weber et al. (2012)
7.	The subunits and components (including details like reliability and maintainability)	*a system's engineering factor*	Teunter and Fortuin (1998), Langseth and Portinale (2007), Weber et al. (2012), Jones, Jenkinson, Yang, and Wang (2010)
8.	The design class	*a system's engineering factor*	Krikke and Van Der Laan (2011), P. Weber et al. (2012)

[3] Also included as 2nd in the list of Section 2.4

SN	Factor identified	Category	References
9.	The life expectancy[4]	*a system's engineering factor*	Ritchie and Wilcox (1977), Hong et al. (2008), Medina Oliva et al. (2009)
10.	The age of the component	mainly *a support factor* (related to repair and replace decisions)	Ritchie and Wilcox (1977), Medina Oliva et al. (2009), (Jouffe, Weber and Munteanu, 2004) Jouffe, Weber and Munteanu (2004)
11.	The planned and the actual preventive maintenance[5]	*a support factor*	Ritchie and Wilcox (1977), Medina Oliva et al. (2009), Jones, Jenkinson, Yang, and Wang (2010), Jouffe, Weber and Munteanu (2004)
12.	The inspection interval	*a support factor*	Medina Oliva et al. (2009), Jones, Jenkinson, Yang, and Wang (2010), Jouffe, Weber and Munteanu (2004)
13.	The success rate of the inspections	*a support factor*	Medina Oliva et al. (2009), Jones, Jenkinson, Yang, and Wang (2010)
14.	The competence of the inspection	*a support factor* (qualitative)	Ritchie and Wilcox (1977), Weber et al. (2012), Jones, Jenkinson, Yang, and Wang (2010), Philippe Weber et al. (2004)
15.	The competence of the maintenance	*a support factor* (qualitative)	Ritchie and Wilcox (1977), Weber et al. (2012), Jones, Jenkinson, Yang, and Wang (2010),

[4] Also included as 6th in the list of Section 2.4

[5] Also included as 3rd in the list of Section 2.4

SN	Factor identified	Category	References
			Jouffe, Weber and Munteanu (2004)
16.	The severity of the failure	*a support factor* (qualitative)	Hong et al. (2008), Medina Oliva et al. (2009)
17.	The costs (spares procurement, holding, etc.)[6]	*a support factor*	Eppen and Iyer, (1997); Hill, (1997); Lariviere and Porteus, (1999). Zheng, Wu, and Shu (2016)
18.	The duration of the period under consideration[7]	*both an operational and a support factor*	Eppen and Iyer, (1997); Hill, (1997); Lariviere and Porteus, (1999). Zheng, Wu, and Shu (2016)
19.	The number of systems (size of the installed base)[8]	*both an operational and a support factor*	Dekker, Pinçe, Zuidwijk, and Jalil, (2013); Jalil, Zuidwijk, Fleischmann, and van Nunen, (2011); Jin and Liao, (2009), Hong et al. (2008), Krikke and Van Der Laan (2011)

The list of factors above that can be used to define the demand context were identified from the literature that was reviewed. However, given that the list has been compiled from many different papers, it was not clear to the author how the factors could potentially interact. Therefore, in order to appreciate the interaction mechanisms and potentially expand the list even more, two Subject Matter Experts (SMEs) were interviewed. Details are provided in Chapter 3.

[6] Also included as 1st in the list of Section 2.4
[7] Also included as 4th in the list of Section 2.4
[8] Also included as 5th in the list of Section 2.4

2.6 Conclusions

This chapter examined the literature on two problems that are similar to the FPP, namely the Newsvendor Problem (NVP) and Last Time Buy (LTB) problem. In the chapter, the similarities and differences of the NVP and the LTB to the FPP were presented. As was discussed, these differences are mainly due to the amount of data that can be available in the FPP as compared to the other two types of logistic problems, and this is also the main reason why there can be other types of forecast models that can be applied to the FPP.

The chapter also identified the factors that have been considered in NVPs and LTB problems to model demand, and that can potentially be used to inform the FPP. Finally, it was suggested that in order to expand on the list of factors and consider how these factors can potentially interact, an additional set of data should be collected, and this is what is presented in Chapter 3.

3 ADDITIONAL FACTORS FOR THE DEFINITION OF THE DEMAND CONTEXT

3.1 Introduction

This chapter aims to identify additional factors that could have an influence on the demand for spares so that they could be included in the models developed later. These factors were elicited during interviews conducted with two SMEs.

Therefore, Section 3.2 presents the interviews conducted to further explore the factors that can contribute to the formulation of the demand context and also to help verify the factors identified in Chapter 2. Section 3.3 presents three conceptual models that can be used to categorise the factors that were identified in the literature and from the interviews. Finally, in Section 3.4 it is argued that even though the demand factors can be identified, the way that they work in order to affect the demand is not easily identifiable by human expertise. However, one's understanding can be enhanced by the exploration of data records of incidents kept in the nodes of the SC.

3.2 Interviews to Further Explore the Demand Context

The contribution of this step to the problem analysis can be summarised in the following graph which highlights the participation in the previously presented Figure 2-1:

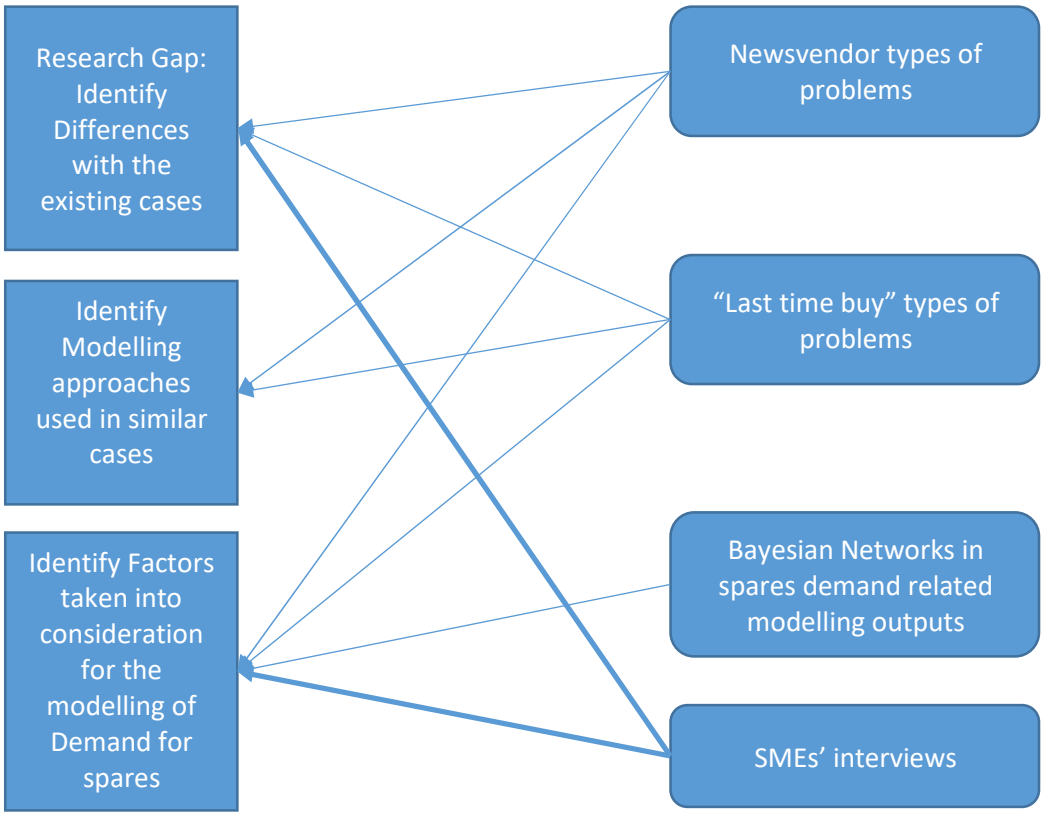


Figure 3-1: The literature review process and objectives updated

The review of factors influencing demand, which was conducted as part of the literature review in Chapter 2, is complemented in this Chapter with primary data elicited from two subject matter experts (SMEs) who shared their knowledge and personal experiences. As stated at the end of Section 2.5, the author was concerned about the level of detail of the factors identified in the reviewed literature, mainly due to differences in the real world applications context of that literature and the present research.

The two SMEs were chosen due to their extensive experience of operational deployments with the British Army and with the RAF. While the number of SMEs was low, few people had the academic and practical knowledge of the specific research topic under investigation. Therefore the lack of experts was to some degree offset by their rare expertise. However, it is acknowledged that the small number was undesirable and constitutes a limitation in this study.

The interview methodology made use of the Critical Decision Method (CDM) (Crandal, Klein, and Hoffman, Robert, 2006), and each interviews took

approximately 60 minutes. One of the core ideas behind CDM is that it uses the critical incident technique during which non-routine, challenging events are probed. The advantages of such an approach are multiple. Since the events are challenging, they call for the specific expertise that the SME can bring to the study. Furthermore, they evoke focused attention and thus important details and causal mechanisms are less likely to be missed.

The second important element of CDM is the gradual deepening on critical points by performing multiple sweeps. The requirement for more than one iteration ("multiple sweeps") is based on the assumption that even though the method helps the interviewee to recall the timeline of an incident, some of the influential details might be missed on the first iteration. It is then up to the subsequent iterations to extend the depth of the exploration for more details at specific points of interest.

The data collection requirement was to understand and record the factors that affect the level of demand for spares in an operation and, of course, the mechanisms between them. In both of the interviews, the first objective was to identify appropriate incidents that were both challenging and relevant to the data collection requirement. In order to achieve that, an imaginative "warm-up" scenario of the logistics support of a number of systems was described. In order to make thinking more realistic, the interviewer defined specific geographical places around the area that the interview was taking place, where each part of the Support Chain (SC) and the operations would take place. The interview then proceeded by describing a challenging incident related to the objective of collecting and mapping spares' demand influential factors. This incident was chosen such that the interviewee observed a gradual ramping-up of the rate of demand in spares related to increased number of breakdowns and the objective of this suggested observation was to focus the interviewee's thinking on the factors related to the demand rate. It was at that point, that the interviewer probed for specific cases from the SMEs' background knowledge by asking the interviewees if they had ever had faced a similar challenging experience.

The interview proceeded by conducting an open-ended discussion on incidents considered relevant to the research. This involved gradually identifying what had actually happened and then mapped the elicited factors and their relations. This process of continued discussion in order to gain ever greater clarification, carried on until potential causal relational factors that could account for the increased breakdown rate of the system, were identified. The resulting set of features are summarised in the following Table 3-1 along with their equivalent as presented in Table 2-1. A descriptive conceptual diagram is presented in Figure 3-2 which was verified by the interviewees.

Table 3-1: Factors that contribute to the demand identified from the interviews

SN	Factors	Category	Source (Literature/Interviews/ Both)
1.	The skills of the maintainer in identifying the correct failure and also performing the repair effectively, and her individual work ethics ("*Maintainers' abilities*" of Figure 3-2)	*a support factor*	Both (similar to SN 15 of Table 2-1)
2.	The applied maintenance policy ("*Repair rate (systems)*" and "*Repair rate (spares)*" of Figure 3-2)	*a support factor*	Both (similar to the SN 11, 12 of Table 2-1)
3.	The effect of the environmental conditions on both the systems and on people. Regarding the people, environmental conditions can affect the willingness of the maintainers to perform the fault identification and repairs to their full needed spectrum, and can also affect the usage choices of the operators	*an environmental factor*	Both (similar to the SN 2 and 3 of Table 2-1)

SN	Factors	Category	Source (Literature/Interviews/ Both)
	("*Environmental conditions*" of Figure 3-2)		
4.	The natural Wear and Tear of the systems ("*Failure rate*" of Figure 3-2)	*a system's engineering factor*	Both (similar to SN 9, 10 of Table 2-1)
5.	The availability of spares ("*Spares*" of Figure 3-2)	*a support factor*	
6.	The placement of wrong orders for spares either in amount or/and types ("*Information distortion*" of Figure 3-2)	*a support factor*	Both (similar to the SN 13 and 14 of Table 2-1)
7.	The effect of cannibalisation practices ("*Cannibalisation*" of Figure 3-2)	*a support factor*	Interviews
8.	The change in operation patterns ("*Type of missions*" of Figure 3-2)	*an operational factor*	Both (similar to SN 4 of Table 2-1)
9.	The skills of the operators and their choices given changes in the operational demands ("*Operators' abilities*" of Figure 3-2)	*an operational factor*	Both (similar to SN 5 of Table 2-1)
10.	The lack of end to end visibility of the SC which translates the occasionally realised delays into lack of trust to the support system ("*Information distortion*" of Figure 3-2)	*a support factor*	Interviews
11.	The geographic location where the systems are operated	*an environmental factor*	Literature (SN 1 of Table 2-1)

SN	Factors	Category	Source (Literature/Interviews/ Both)
12.	The system's code / particular configuration of the system	*a system's engineering factor*	Literature (SN 6 of Table 2-1)
13.	The subunits and components (including details like reliability and maintainability)	*a system's engineering factor*	Literature (SN 7 of Table 2-1)
14.	The design class	*a system's engineering factor*	Literature (SN 8 of Table 2-1)
15.	The severity of the failure	*a support factor* (qualitative)	Literature (SN 16 of Table 2-1)
16.	The costs (spares procurement, holding, etc.)[9]	*a support factor*	Literature (SN 17 of Table 2-1)
17.	The duration of the period under consideration[10]	*both an operational and a support factor*	Literature (SN 18 of Table 2-1)
18.	The number of systems (size of the installed base)[11]	*both an operational and a support factor*	Literature (SN 19 of Table 2-1)

[9] Also included as 1st in the list of Section 2.4
[10] Also included as 4th in the list of Section 2.4
[11] Also included as 5th in the list of Section 2.4

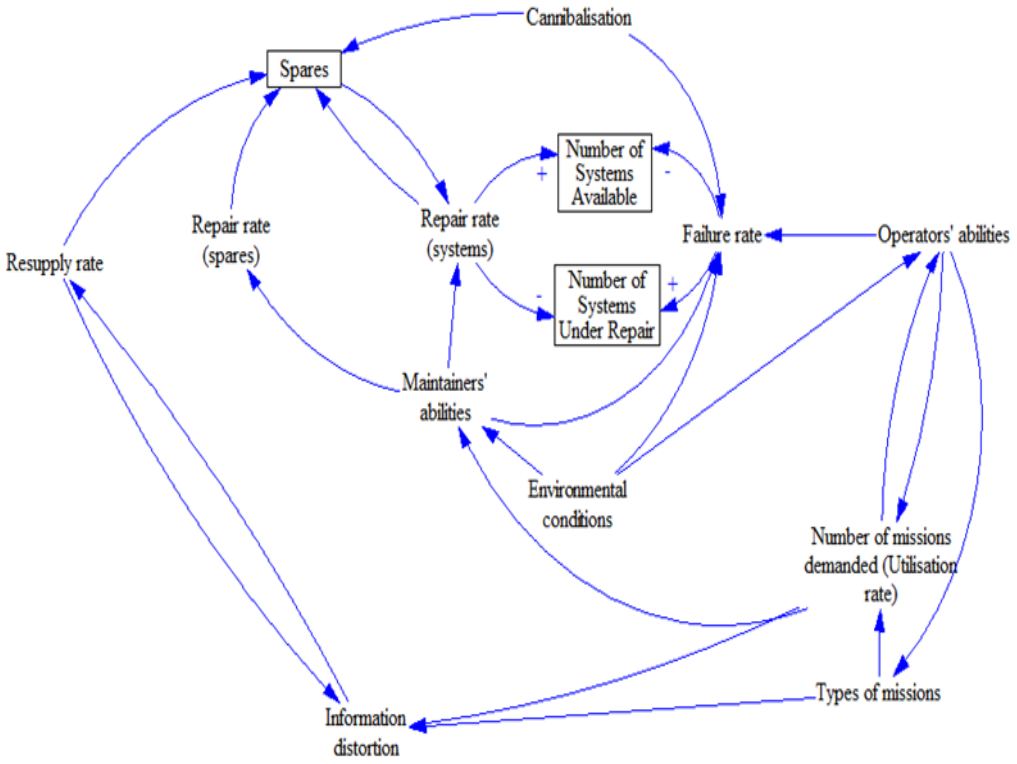


Figure 3-2: Causal factors and their relations as elicited from the interviewed SMEs

The interviews verified most of the factors identified from the literature and also the fact that the demand is affected by activities propagating in the SC. Furthermore, the interviews also highlighted additional factors that cannot be easily measured or estimated. One such factor is when the decision makers have a fixed but incorrect view about the root cause of the problem For example a decision maker firmly believes that the cause of the experienced low fleet availability is the lack of spares, while the true root cause is the lack of adequate experience in fault identification which causes the demand and consumption of the wrong spares. Other factors that were identified from the interviews were that the environmental conditions do not only have an effect on the systems but also on the operators and on provision of support, and also that the lack of trust in the supply chain affects the resupply rate.

The identified lack of trust in the supply chain as a factor that can have an influential effect on the demand for spares is a common phenomenon

(Christopher and Peck, 2004; Chu, Chang and Huang, 2011). As elicited by the interviewed SMEs and also suggested by Christopher and Peck, the lack of trust can be caused by the lack of visibility in the supply chain which is a cause of uncertainty and occasionally of risk. Consequently, this also suggests that there is no single expert who can have a complete understanding of the Supply Chain, not to say the whole Support Chain (SC) and its interactions with the Operations and the Environment. This observation along with the fact that access to expertise is not easy either, suggests that the knowledge needed in order to build a model of the demand's context would probably not be complete if it relies only on the expertise of those in the SC. What the present research suggests is that the knowledge gap can potentially be reduced by related SC data accompanied by the resulting model(s) of the demand context, and this is something that has also been demonstrated in the findings of Chapter 7 and in Appendix B.

Even though the literature review accompanied by the analysis from the SMEs' interviews revealed additional factors to be included in the demand context along with a suggested way in which they can interact to formulate a context, at the same time an additional question has also been raised.

This question concerns how the factors can be identified in any similar specific case. There can be problems in which the modeller might need either to verify or even formulate her own view of which factors to look at. Consequently, the requirement is to be able to use the factors' identified categories in order to build conceptual models that can help in the identification of the spares' demand formulating factors. The next section discusses three such conceptual tools.

3.3 Conceptual Tools for the Identification of the Demand Context Factors

In what follows three conceptual models are presented that were used in either categorising existing factors or identifying others that could potentially be considered for inclusion in the demand context that is to be modelled.

The first conceptual model is related to the function of the Operational Availability metric. A careful look at the whole set of factors as presented at the end of Sections 2.5 and 3.2 and that resulted from the review of the literature and from

the interviews of the SMEs, revealed that they are directly related to a function that is commonly used to express the long-run Operational Availability A_o metric (Pryor, 2008):

$$A_o = \frac{MTBM}{MTBM + MTTR + MLADT}$$

$MTBM$: Mean Time Between Maintenance activities (either corrective or preventive)

$MTTR$: Mean Time To Repair

$MLADT$: Mean Logistics and Administrative Delay Time

So, what is apparent from the above, is that the *operational* and the *environmental* factors have a direct effect on the repair frequency and thus on the $MTBM$, and so do the *systems engineering* factors which also affect the $MTTR$, while the *support* factors affect the $MLADT$.

Therefore, in order to identify potential factors that can constitute the demand context using this model, the analyst can use the following questions:

1. What defines the value of the time between maintenance (corrective/preventive)? What can cause the time between maintenance to increase? What can cause the time between maintenance to decrease?
2. What defines the value of the time to repair? What can cause the time to repair to increase? What can cause the time to repair to decrease?
3. What defines the logistic delay / administrative delay times? What can cause the increase of the logistic delay / administrative delay times? What can cause the decrease of the logistic delay / administrative delay times?

However, as can be seen from the above list of questions, this conceptual model prompts more emphasis on the support activities as presented in the lists at the end of Sections 2.5 and 3.2, than on the others.

A second conceptual model that is related to the previous one, comes from the observation that the Operational Availability (A_o) can be considered as a measurement of the outputs from the interactions among the support and operational activities of a specific system. Therefore, the factors listed at the end

of Sections 2.5 and 3.2 can also be visualised as resulting from the (dynamic) interactions among the nodes of a Support Chain (SC) up to and including the operations that are supported, and thus a process diagram of the SC activities can work as a conceptual model for the elicitation of these factors (Figure 3-3).

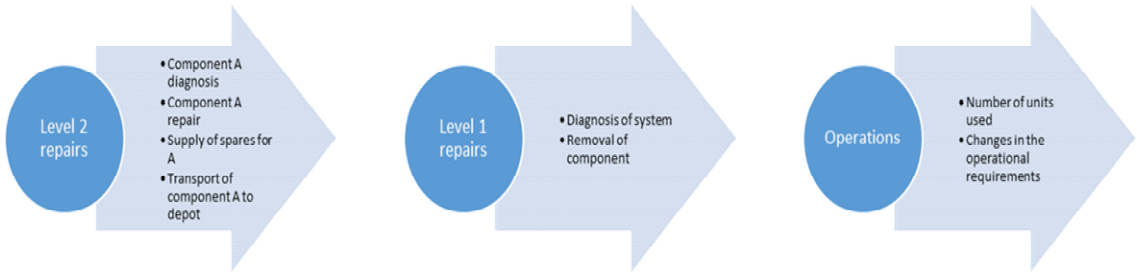


Figure 3-3: Example of a process diagram that can be used as a conceptual model for the identification of the demand context factors

In the example of Figure 3-3, the demand for spares is experienced at Level 1 of the SC. Given the local and network activities, the analyst can identify candidate factors like those related with the 'Diagnose' activity (quality, rate, capacity, etc.) at the Level 1 node, or those related with the 'Changes in operational requirements' activity that originates from the Operations node.

The third conceptual model comes from some earlier observations. As the categorisation of the factors presented at the end of Sections 2.5 and 3.2, strongly suggests that the condition of the systems that are operated and supported, is affected by four interacting contexts. The first context is the engineering system to which they belong. This concerns their reliability, their engineering structure and their maintainability. The second context is the operational one. This is the context within which the systems are used and have their components "worn-out". Consequently, it is the factors that define the operational context that cause the reduction in the pool of the deployed systems. On the other hand, the third context is the one that is responsible for replenishment of the pool of the deployed systems. This is the support context in which the maintenance and logistics/supply activities take place. The final context is the environmental one, into which both the operational and support contexts are embedded, and which affects the deployed systems, either during

their operating or their stand-by time. In summary, the four contexts of this third conceptual model are as follows:

1. Engineering system
2. Operational context
3. Support context
4. Environment

Using any of the three conceptual models, or their combination, the modeller can categorise the factors that have been identified in the present research (Table 2-1 and Table 3-1), or even identify them in a different relevant study.

This thesis mostly used the third conceptual model for the development of the simulation (Sections 6.3.1, 6.3.2, 6.3.3, 6.3.4). The reason is that this conceptual model seems to be closer to the way that the factors were provided by the interviewees during the interviews.

3.4 Data as an Important Supplement to Experts' Knowledge

Studying how the factors interact not only facilitates an understanding of how demand is affected by them, but also drives the development of the demand forecast models. Consequently, the ability to identify how the factors affect each other and the demand is important.

In the literature review on NVPs and the LTB/EOL problems, it was noted that according to the type of business that was assumed, the models focused on only one of the participants in the SC, namely either the "buyer" or the "manufacturer" / "service provider". In such cases the partiality of the view of the influence of the SC activities on the experienced demand is more or less inevitable, and this influence is probably the reason that the models that have been applied have to rely on assumptions of the anticipated repair choices, aggregate data, and consider long time periods, etc. On the other hand, the FPP cases examined with the SMEs, even though they were referring to a wider view of the SC due to the closer relationships among the SC participants, revealed that it is not easy for any single SME to have a holistic view of the SC (SN 10 in Table 3-1).

Nevertheless, such an ability would facilitate a fuller understanding of how the factors interact to formulate the demand context.

However, where relationships in an SC are close, there is also an attribute that can potentially be used to increase the understanding of how the factors interact to affect the demand, and in this way help in the formulation of the demand model. The idea is that the effects of the interactions result in incidents which are recorded in the logbooks of the different nodes of the SC, and given that the relationships among the SC participants are assumed to be close, the access to such data is permeable. So, if, for example, the lack of trust in the SC causes an increase in the inventory supply orders due to more frequent and/or larger orders, the on-hand inventory will be affected as well. The latter will be recorded in the logbooks of one of the SC's depots. Furthermore, this can also have an effect on the repair activities which will be recorded in the repair shops.

The data records kept in the logbooks of the SC nodes can include a number of factors of interest. However, the most interesting observation is that these records can be from different nodes and this fact suggests that the interactions among the factors at work in the different nodes of the SC can potentially be captured in the datasets.

Regarding the qualitative factors, some of them can potentially be inferred by the records of relative indicators. So, for example the skills of the maintainers that take over a repair job can be potentially inferred by their years of experience or their rank. On the other hand, some other qualitative factors, like the quality of higher management decisions, are not likely to have been captured, but their effects probably will have.

3.5 Conclusions

Chapter 3 further explored the possible factors that can contribute to the formulation of the demand context. Using interviews, a number of the factors already identified in Chapter 2 were cross-validated and a few more were elicited. Furthermore, three conceptual models were suggested that can be used to prompt thinking when in a specific case a modeller wishes to facilitate the

identification of factors that can affect the demand for spares. One of these will be used in Chapter 6 to help develop the simulation model of the support system.

The Chapter also argues (Section 3.4) that a demand forecast model cannot be developed using only the understanding of the SMEs on how the factors interact, since that understanding can only be partial. Using this argument it goes on to suggest that a large number of the interactions required to build the model can be elicited from the records kept in the logbooks at the nodes of an SC. Thus the data included there can complement the SMEs' understanding of the factors' interactions and in this way facilitate the development of a demand forecast model.

4 METHODS

4.1 Introduction

This chapter presents the methods used in the thesis to model the forecasts of demand. Firstly, the methods of developing the BNs are presented followed by methods for the discretisation of continuous variables which is required as a data preparation step. Furthermore, two other modelling approaches used to provide comparative forecasts are presented, namely logistic regression and the SME's judgmental adjustment of a simple exponential smoothing model's forecasts. These two approaches are chosen because of their wide popularity in this domain.

4.2 Bayesian Networks (BNs)

4.2.1 Characteristics of the BN Models

BNs belong to the family of probabilistic graphical models which is a class of models that use graphs for the representation of probabilistic relationships among the variables of interest (Jensen and Nielsen, 2007a; Madigan, York, and Allard, 1995; Pearl, 1988b). Graphical models in general and BNs in particular have a wide spectrum of applications due to their flexibility and interpretability (Hartemink, 2001).

A BN is a Directed Acyclic Graph (DAG) in which there are nodes/variables that represent the variables of interest. The nodes can be connected to each other with an arc and the node that is at the head of the arc is called the child node, while the node at the other end is called the parent node. A child node can be connected with more than one parent node and the reverse (several parent nodes to a single child node). An arc that connects nodes X and Y encodes the assumption that there is a direct association between the two which can qualitatively denote a causal or influential link between the two, while the DAG is called "Acyclic" because the arcs should never create a cyclical path. Each node has a set of conditional probability values associated with it, formed in Node Probability Tables (NPT) or else Conditional Probability tables (CPT) which model the uncertainty in the relationship between the node and its parents. Of course, if

a node has no parents, its NPT is the probability distribution of its values, and. such nodes are called root nodes. The efficiency of the BN's structure is mainly owed to its graphical properties.

4.3 Bayesian Network Structure and Node Probability Tables

4.3.1 Building the Bayesian Network (BN)

The building or learning of a Bayesian Network includes two processes:

1. Structure learning, in which the model graph is built
2. Parameter learning in which the local probability distributions are learnt based on the chosen structure

These two processes are performed by either feeding data into a learning algorithm (data-based approach), or by eliciting the associative relationships among variables from SMEs (knowledge-based approach), or a combination of the two. In the first method, the variables used are only the ones that the modeller has data on, while in the second, variables that might not be included in a data set – latent variables - can be elicited from SMEs and used to describe the context of interest better. However, data-based approaches can reveal relationships among variables that are not easy to get from SME's. This is what the interviews presented in Section 3.2 also indicated. Especially when the within-scope system is extended and the variables involved are extended in space, eliciting an adequately valid BN from experts can be very challenging (Scutari and Denis, 2015b). On the other hand, it is a common procedure to use a combination of methods (Hartemink, 2001; Heckerman, Geiger and Chickering, 1995).

In the cases that the present research examines, the data are from the past periods of the building up of operations and of their infinite-time horizon phase. However, since the decisions are for the final phase of the operations (FPP) and these decisions can be very challenging due to the number of changes that are planned to take place and the uncertainty involved with the effect of those changes on the systems' failures and the resulting demand for spares, the contribution of SMEs' expertise is also important. Therefore, in the specific nature of the problem, for the development of the BN, the modeller would probably need

not to rely just on past data but to also seek for verification and at least face-validation from SMEs.

4.3.2 Structure Learning

In the structure learning the modeller tries to identify the graphical structure (the DAG) of the BN. The objective is to find for each node the minimum set of connections around it that correspond to its associations with the rest of the nodes.

Structure learning algorithms from data are mainly classified in three categories: constraint based, score based and hybrid. The assumptions under which the algorithms operate are the following (Nagarajan, Scutari and Lebre, 2013):

1. Every node in a DAG represents a single variable and every variable is represented by a single node. This means that resulting nodes must not be functions of another. This assumption is needed for the unsupervised learning of the DAG. Furthermore, the assumption does not exclude after the BN has been built by the learning algorithm, amending or expanding it by including nodes and arrows that model deterministic relations (see e.g. the "definitional" idiom in Fenton and Neil, 2013, or Neil, Fenton, and Nielson, 2000)
2. The building blocks of a BN are the conditional independencies and therefore all relationships calculated between the BN's variables are seen as such
3. All observations are independent realisations. If there is some known kind of dependency (e.g. spatial or temporal through a latent variable), then it must be accounted for (Cooper and Herskovits, 1992; Heckerman et al., 1995)
4. The existing combination of the values of all the variables under consideration must have a non-zero probability. If not, then the Markov Blankets cannot be uniquely identified and neither the BN model

4.3.2.1 Constraint Based Algorithms

This class of algorithms are called constraint-based since, in order to reconstruct the BN structure the algorithms are driven/constrained by the existing conditional independence relationships that exist in the data and expressed among each one variable and the rest within the domain. Other types of constraints might be used as well which are not related to conditional independence, e.g. in cases where there might be latent variables. The present research does not refer to these types of constraints. Furthermore, there are algorithms that instead of conditional independence, use mutual information tests and are tested over the asymptotic or semiparametric χ2 distribution, sequential Monte Carlo permutations etc. (Scutari and Denis, 2015a)

Several conditional independence tests between pairs of variables conditioned on a set of others are performed to guide the gradual positioning of the nodes and their connections in the network. The null hypothesis H_0 of the performed tests is that the examined variable and a set of others are independent. This set can also be the empty set ∅.

The quality of the constraint-based algorithms depends on the efficiency in the formulation of the sets that are tested for each variable and the reliability of the tests. Both of these issues depend on the relative size of the variables' domain as compared to the size of the available data set (Dash and Druzdzel, 2002). In the cases examined, there can be a plethora of data coming from the logbooks. Nevertheless, the size of the variable domain can be large as well. As shown in the literature review and reinforced by the findings from the interviews with the SMEs, the demand context is formulated by the Environmental, Engineering, Operational and Support contexts while it extends to the different nodes of the SC and operations. Choosing which variables to include in a model is very important in the development of a model and of its ability either to predict or to explain. Furthermore, the choice of the variables is also directly related to the method to be used in order to build the model (Field, Miles, and Field, 2012).

For the reasons discussed in Section 4.3.2.3, the datasets used in the present research did not favour the application of constraint based algorithms. On the

other hand, there was no particular problem in applying the score based algorithms that are presented next (Section 4.3.2.2)

4.3.2.2 Score Based Algorithms

The score based algorithms do not rely on independence tests among combinations of variables, but on evaluating a candidate hypothesised BN DAG as a whole. Score based algorithms seek to find the most probable network structure given the available data (Cooper and Herskovits, 1992) and it is the resulting structure as a whole which provides insight to the dependency relationships among the variables. In their general approach, by thinking of the search for the most appropriate BN structure as an optimization problem, the possible relations among the variables/nodes as a state space can be conceptualised. Therefore, score based algorithms need:

- A state space,
- An initial BN structure,
- A termination condition evaluated over a metric that expresses the fit of the BN on the provided data sample D and
- A search engine that efficiently iterates among different candidate structures within the state space

The assumptions adopted for the development and use of the Bayesian Dirichlet equivalent (BDe) score are the following (Cooper and Herskovits, 1992; Gilks, Thomas and Spiegelhalter, 1993; Heckerman, Geiger and Chickering, 1995):

Assumption 1: The values of the variables have come from a *multinomial* distribution. This is a reasonable assumption for many of the variables for datasets coming from FPPs, since the data are sourced from the logbooks which record incidents, like a component's breakdown, a diagnosis completion incident etc. However, this assumption excludes other variables which are numeric and infinite. In the cases examined in the present research, such a variable can be the "Number of hours that the component X has been operating". For these variables, a discretisation pre-processing step is required.

Assumption 2: The variables' multinomial distributions parameters are independent in the network structure (*global parameter independence assumption*) and also the parameters of a variable associated with the different states of its parents are independent (*local parameter independence*).

Assumption 2 follows from assumption 1 and combined with assumption 3 and 4 aim to simplify the computations for the score metric. In Section 4.4.1, a discussion is presented on the benefits and challenges for the FPP of transforming the variables to multinomial. However, for the present assumption the anticipated benefit is to get computational efficiency that leads to a single, additive metric.

Assumption 3: In two different possible network structures, if a node has the same parents in both, then the NPTs will be the same. This is called *parameter modularity assumption*.

Assumption 4: In a possible network structure the values of the probabilities of any variable, yet unknown, follow a *Dirichlet distribution.*

As mentioned above, the assumptions 2, 3 and 4 have been introduced to help with the computational efficiency of the metric.

Assumption 5: The dataset D is assumed to be complete. If D is not complete, the algorithm that counts the elementary events in the variables to estimate the probabilities, will not be able to work.

It should be expected that this would be a challenging assumption when real life data are acquired. Occasionally, there are missing or wrong data points, which though can be handled with statistical analysis. However, for the present thesis where the data were acquired from a simulation, this problem did not exist.

Assumption 6: Given two network structures B_{S1} and B_{S2} which are both valid DAGs (i.e. no loops, etc.), if they are equivalent (Chickering 1995), then the likelihood that the dataset D has come from B_{S1} is equal to the likelihood that it has come from B_{S2}. Therefore, given this assumption, the distribution of their parameters is the same as well. This assumption is called *likelihood equivalence* and it gives the "e" for equivalence to the BDe metric.

Assumption 6 facilitates the search algorithm by reducing the search space. Such a reduction is computationally very important, especially in the FPPs where the number of variables can be high.

Furthermore, in order to facilitate the calculations for the prior distributions of the hypothesised BN structures, Heckerman et al. (1995) adopt Buntine's (1991) suggestion to assign an equal probability to every state in the domain and to every possible structure. This is a special case of BDe and is called BDeu ("u" for "uniform"). Actually, BDeu is the only member of the BDe scores that is in common use (Scutari and Denis, 2015a).

Regarding the prior distributions of the hypothesised BN structures, they can be any structure that can be elicited by a SME, or even the output of a constraint-based learning algorithm as is the case in the hybrid algorithms (Scutari and Denis, 2015b). These hybrid algorithms are not to be confused with the "Hybrid" BNs which refer to the continuous and discrete types of variables that they can incorporate. Moreover, for the purposes of the present research, the approach of building the DAG structure through the combination of an expert-elicited BN structure and then applying machine learning is also called hybrid (Sections 7.2.3.3 and 7.2.3.4).

Another usual prior structure that can be used as a starting network for the optimization algorithms is to use a random prior structure. Using multiple such random starting structures helps in covering the search space more thoroughly and not including any systematic bias. A relative algorithm of random starting structures is the one proposed in Ide and Cozman (2002) and applied in the bnlearn R-package (Nagarajan, Scutari and Lebre, 2013). This specific algorithm is the one that has been used in the present research (Section 7.2.3.1).

Optimisation algorithms converge to local optimum solutions. These resulting solutions depend not only on the algorithm itself, but also on the starting point. Consequently, it is common practice to store the many local optima which have been generated by running the algorithms a number of times (arbitrary chosen to 300) and each time starting from a different initial structure (an initial B_S^h) that has been randomly created. These many local optima have a number of their arrows

in common. The practice then is to keep that “averaged” structure that has those arrows that are in common to the majority of the many created optimal structures, with the majority cut-off value defined by the user.

4.3.2.3 Evaluation of the Applicability of the Constraint-Based and Score Based Algorithms in the Final Phase Cases

When any constraint-based / local learning algorithms was applied to the dataset that was used for the present research, it was observed that the many of the expected dependence relations did not occur (see also Sections 7.2.1 and 7.3.1). The resulting graphs had very few nodes connected, while many of the nodes were presented as not being associated to any of the rest. The conclusion drawn about the reason for this result has to do with the peculiarities of certain key variables in the datasets.

The datasets that have been used in the present research, include those variables that define the context of the demand for spares. Within this set of variable a key role is played by those that capture the failure incidents of the components (e.g. FRT_LRU, FRT_PRU or FRT_DU in Table 6-1). Moreover, real-life supported systems are built in such a way that their components are very reliable and consequently their probability of failure is engineered to be very low (Sherbrooke, 2004, p.6). This means that in the recorded values of the failure incidents only very few cases were failures while the rest were non-failure incidents (for a two-state variable). Consequently, the information which is of key value – the failure incident – is rare within the dataset, a fact that can be problematic for the independence tests applied in the constraint-based learning: rare events can result in (falsely) not rejecting the null hypothesis H_0 that the variables are independent (a Type II error) and thus not introducing (or, depending on the algorithm, not retaining) the edge between the tested nodes.

Spirtes et al. (2000, p. 96) made a very relevant observation: local learning algorithms might suffer from the fact that after falsely excluding a connection between two nodes then this can result in further multiple false disconnections. As a mediation, Spirtes et al suggested to use a Bayesian procedure like the one

presented by Cooper and Herskovits (1992) as a repair step to the constraint-based output structure.

The above described limitation of the constraint based algorithms is not a problem for the score based. Moreover, the latter were chosen for this research because they offer an additional modelling benefit. Score based algorithms have a very useful virtue inherent to their scoring metric (Hartemink, 2001, secs4.3, 4.4). The metric takes an average over a family of probability distributions which works as a penalty for unnecessary parameter complexity. This fact, as compared to alternative scores that use a single maximum a posteriori parameter, is an inherent guard against parameter overfitting especially when the available data are comparatively few and it is also beneficial when faced with noisy data. Furthermore, the scoring metric has also got a provision to permit the inclusion of prior experts' knowledge.

However, there is still the issue that the structure development method needs to use optimisation algorithms that identify multiple different local optima of the score and thus need to be run many times. This practice can create model structure overfitting. In order to overcome this issue, it is not the single optimum structure that is retained but rather it is the "average" over all the structures so that only those arrows are retained which appear in above a predefined percentage number of structures (Nagarajan, Scutari and Lebre, 2013).

4.3.2.4 Causal and Acausal BNs

Hartemink (2001) and Heckerman et al. (1995) make an important observation about the structure learning algorithms and their interpretability. In the authors' explicit assumptions for the development of the BDe score metric, they make a clear dinstiction between the "causal" and the "acausal" structures.

In more detail, in order to simplify the calculations of the BD metric by constraining it into the BDe the authors adopt the equivalance hypothesis. The hypothesis states that the srtucture B_S^h is true *iff* (if and only if) the database is a sample of multinomial variables that have resulted from B_S. This hypothesis is satisfied *iff* the resulting parameters Θ_U satisfy the conditional independencies of the true structure B_S. The direct consequence is that if two not the same structures are

equivalent, then their hypothesised structures can be equal. This means that if, for example, the changing of the direction of an arrow produces the same score, then the two different hypothesised structures are equal. However, such an assertion could violate any understanding of causality that might be infered by the arrow's direction.

Using a similar observation, Hartemink (2001) suggests that the BN user should be careful about the structures' interpretation. The author correctly highlights the difference between statistical interpretations and causality. In the cases of the present research, many different demand context mechanisms might map to the same set of statistical dependencies. Additionally, in the core of the present research objectives is to be able to build BNs by the use of the data recorded in the logbooks of the nodes of the SC and of the operations. In such cases it is logical that the system has not been observed by the same humans in a number of different configurations and thus certain causal dependencies might not have been apparent. This is something that the present research's use of simulation to produce replications of multiple possible futures has been able to reveal (Appendix B). A further direct consequence of the low breadth of the SC and operations' observations in different settings is the possible existence of a number of latent variables that can have a confounding effect on the modelled mechanisms, while either the existence or the values of these variables might not be recorded or known.

To cope with the interpretability of the BNs' structures, advocates of the preservation of their causal character/merits (Fenton and Neil, 2013, sec. 7.2; Neil, Fenton, and Nielson, 2000) call the arrows "causal or influential", thus recognising that the complete causal relations among the variables might not be fully known. Nevertheless, it is this same notion of causality and influence that helps in identifying the structure of the BN from the knowledge of SMEs.

4.3.2.5 Eliciting the Structure from SMEs

As mentioned earlier, a prime objective of this research has been to be able to use the records kept in the logbooks of different nodes of an SC and the supported operations. Driven by this objective and the related extended breadth

of variables involved, the development of the corresponding BN models can mainly be through machine-learning algorithms as discussed earlier (Section 3.4). However, prior network structures, and even more importantly variables like the “Environment” or the “Type of Operations” that might not be included in the logbooks, but can have an influential effect on variables like the “Rate of Use” and “Failure Rate”, should be considered even if data are not readily available.

The same methods that have been used in the constraint-based algorithms, i.e. d-connection/d-separation, can be used for the elicitation of the BN structure. Therefore, serial, diverging and converging connections can be used to build topologies of small numbers of nodes by thinking what the effects of entering evidence in one node is to the propagation of evidence between the other two and then connect them in a bottom up manner. Briefly in a set of three nodes A, B and C:

- If information about node B renders any new information about A not affecting the belief about C then either a serial or a diverging connection. However, if a new information about A can still affect the belief about C even if information about B is $available$, then the simple serial or diverging connections are not appropriate. Under these circumstances consideration should be given to whether is a need to connect A to C or to use a different type of connection
- If the evidence between A and C can only be propagated when there is evidence on B, then a converging connection exists. However, if there is the understanding that information on A can influence C even if there is no evidence about B then the simple converging connection might not be enough. In such circumstances consideration should be given to the possible need to have a direct link between A and C or a different type of connection

However, the above approach is difficult to implement by experts who are not familiar with thinking in terms of conditional dependences. Laskey and Mahoney (1998) recognised the need to create “fragments” of networks. According to the authors, each fragment is a grouping of nodes that are related to each other and

that can be thought about in isolation from the rest. The grouping is suggested by the SMEs and they need to provide some underlying reason for the nodes to be considered together. However, for the purpose of building a complete BN structure Laskey and Mahoney's (1998) suggestion is still not detailed enough and a more detailed process/method is required. Typical practical problems that need to be addressed are (Neil, Fenton and Nielson, 2000):

- What is the direction of the edge - if any - between two nodes that best describes their relation?
- How much detail is needed in the identification of the nodes?
- How can the structure be managed so that the number of parents in a node are kept small?
- How can experience be codified and reused in other problem cases?

These questions led to an advancement in the methods that can be used in order to elicit expert knowledge for the structure building of the BNs. The method of elicitation discussed below was introduced in Neil, Fenton, and Nielson (2000) (see also Fenton and Neil (2013)). It includes a set of abstract patterns that Neil, Fenton, and Nielson call “idioms” which can be used as building blocks of the BNs’ structures as described by the SMEs. The idioms have been developed through the identification of common patterns in the development of BN models from SMEs. One of the fundamental attributes of these idioms, which is also core to the value-adding use of BN modes, is that they are built in such a way so that the resulting structural components can be explained to and verified by domain experts. Furthermore, as expected these idioms use and preserve the d-separation/connection properties that are needed for a BN. The modeller along with the SMEs define fragments of the variables and related idioms, and the modeller can turn the idioms in BN objects that are gradually integrated in a bottom-up manner to create the model.

The sections that follow present those idioms that were used in the present research to elicit the relationships among the variables.

4.3.2.5.1 The Cause-Consequence Idiom

The cause-consequence idiom is used to model a causal process. The minimum set that can be included in this idiom is two nodes with an arrow. As expected the node at the tail represents the event or facts that the process needs as an input (the cause), while the event or factor that is the output of the process (consequence) is at the head of the arrow. Care should be given that it is the arrow that models the process itself and not the nodes. The process can be thought of and is modelled via the NPTs of the consequence node. Furthermore, the cause/input of the process can be a transformation of the same input or a new output (Fenton and Neil, 2013, p.176).

There are also process categories that can facilitate the identification of cause-consequence relations:

- Productive/mechanical category. Production plants can naturally be included in this category. Furthermore, more abstract but related problems can be included, like the quality of a design which is causally connected to the number of failures, the complexity of a problem to the number of failures, the skill of a mechanic to the time to repair a fault, etc.
- Physical/natural category. Examples in this category can be the environment's effect on the wear-out of a component, or its effect on the duration of a transportation
- Intentional category. Examples in this category are those in which there is an intention to incur an output, like the event of hacking an ICT system and its outcome

Another distinguishing attribute of this idiom is that the cause and the consequence events/facts have a sequence; a chronological order. This creates a challenge to the modelling of datasets like the ones dealt with in this research. The datasets that the present research assumes access to are records from different logbooks. Consequently, each recorded value of the variables - each individual case - refers to the same time-instance. This is an important point because it drives how the modeller and the SME who might be helping in the development of the BN structure should be understanding the relationships among the variables and therefore, the specific idioms that can be used. The fact

that each case in the dataset refers to the same time-instance, means that if for example a failure event took place and at the same time that the weather was bad, it cannot have a cause-consequence relationship as intuitionally might be considered to be the case. The two incidents do not have a chronological order. Nevertheless, the cause of the failure instance is not the bad weather instance of that particular time but the accumulated bad weather instances before that specific time-instance (and of the similar effects of other causal factors).

On the other hand, there are variables' relations which can still use this idiom even by the application of the logbooks as the only source of data. An example is the skill level of a mechanic and the repair output. In a repair shop's logbook, a repair job is recorded and at the same record/time-line, it is allocated as a task to a specific mechanic who has a certain repair skill-level (and taking into consideration the interviews with the SMEs, mechanics can also have different work-ethic levels as well (SN 1 in Table 3-1)). It is reasonable for a highly skilled mechanic to be more productive and thus have the job for a shorter period of time. Consequently, the skill level has a causal effect to the repair output which is also captured in the logbooks.

Nevertheless, the earlier example of the relationship between the environment and the failure incidents can be dealt with the following idiom of measurements/indicators.

4.3.2.5.2 The Measurement/Indicators Idiom

In the cause-consequence idiom, the two nodes represented two different attributes. On the other hand, there are a large number of cases which have estimates/judgements/indicators of a single variable. Therefore, it is necessary to eventually have to consider the estimates and the true value of the variable. However, in such cases a third variable is involved which expresses the uncertainty in the accuracy of the estimate, which can also expand to include biases or intervening circumstances. A generic representation of the measurements/indicators idiom is presented in Figure 4-1.

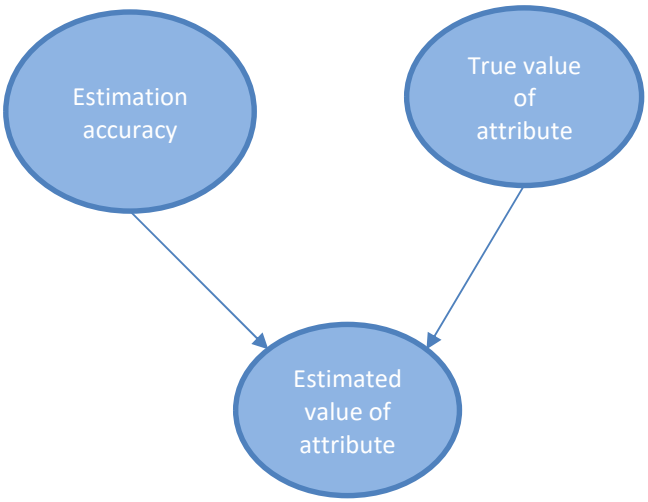


Figure 4-1: Generic measurement/indicator idiom

Through the introduction of the accuracy node, the variables' relationship that is modelled by the idiom develops a very valuable characteristic i.e. to be able to reason by explaining away false positive results.

Another way to see the above triplet of nodes is that the output (what is actually measured) is a combination of the intensity/mass of the true value and of the quality/accuracy of the measuring process. These two nodes of intensity/mass/true value and of quality/accuracy exist before the measurement, so in this sense there is still a causal relationship indicated by the arrows.

However, it must be pointed out that there is not always a need to have a separate node for the "accuracy" of the measurements. For example, the *number of hours* that a component has worked without maintenance can be considered as an indicator of its tendency to fail. In such a case there might be a tendency to model their relationship using the cause/consequence idiom by placing the *number of hours* node at the tail of the arrow and the *failure* node (with values "Yes" and "No") at its head. However, the relationship of the *number of hours* and the *failure* is not that of a process or some kind of a transformation that takes the *number of hours* as an input and turns them to a *failure* as an output/product. The relationship between these two variables might not be thought of as a process, but rather be modelled using the high number of hours to work as an "alert"/indicator for the tendency of the component to fail. Furthermore, the

strength of this relationship can be captured by the NPT of the child node (the hours worked) without the need for an additional node that captures the accuracy of the measurement/indicator.

The earlier discussed situation between the recorded values of the environmental conditions variable and the simultaneously recorded value of the state of a component ("Working" or "Broken down") (Section 4.3.2.5.1) can be considered in a similar way and thus use the indicator idiom. In a narrative form, the presence of a harsh environment is an indicator of the tendency to experience malfunctions in certain components, while the NPTs capture the strength of such a relationship without the need of an accuracy node to be present.

The following idioms do not have to do with the consideration of cause and effect but rather formalise modelling practices that expand the areas of application of the BNs.

4.3.2.5.3 The Induction Idiom

The induction idiom's arrows directions do not indicate causality, even though of course causal links can be set afterwards with other idioms of cause/consequence or measurement/indicators. The induction idiom is a very useful expansion of the BNs' applications in calculating and representing statistical inferences. Through the BNs, a population's parameter is estimated using the available data and this parameter can then be used in a statistical model, including of course another BN's node that uses a cause/consequence and/or measurement/indicators idiom

The way that the induction idiom can be applied is demonstrated in Figure 4-2.

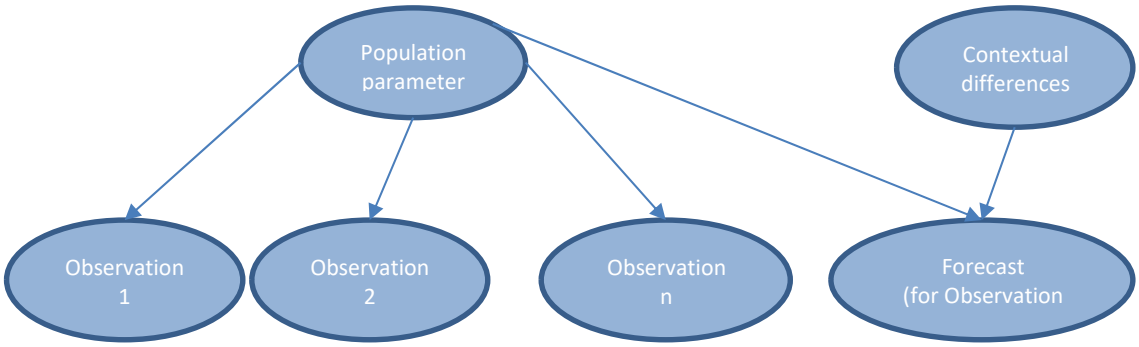


Figure 4-2: Generic induction idiom using data

Assuming that a dataset of n exchangeable observations, then it is possible to use a BN structure like the one in Figure 4-2 to make a forecast about their $n+1$ value. The BN section with the population parameter node is given an initial prior distribution and then is updated by the use of the n observations. The updated population parameter distribution is then used to provide an estimate for the $n+1$ (not yet acquired) observation. Furthermore, it is possible to adjust the forecast on the $n+1$ prediction, which might not be exchangeable to the earlier, by the inclusion of any potential knowledge available from using the contextual differences node.

An additional benefit that the BNs can provide as compared to classical statistical inference models is in case of the absence of data. In such cases, the distribution parameter can be elicited from SMEs and used as shown in Figure 4-3. The lack of data is compensated by the elicited knowledge while again any difference between the context of the past knowledge and the future context can be modelled through the contextual differences node.

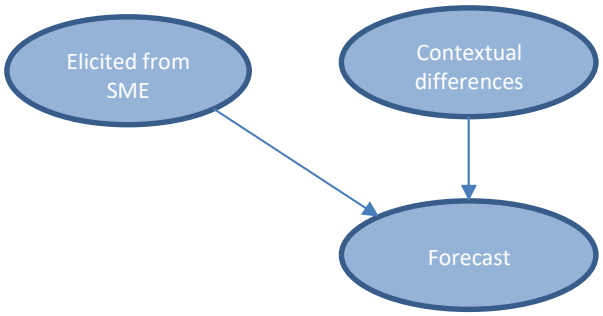


Figure 4-3: Generic induction idiom without data

In practice, there can be more than one population parameters to be learnt and updated as in the basic model for Bayesian inference.

4.3.3 BN Node Probability Tables (NPT)

The step that follows the development of the BN structure is the estimation of the NPTs that describe the conditional independence relationships among the nodes as these relationships are expressed in the structure. In the cases that the present research refers to, it is expected that most of the conditional probabilities are able to be calculated from the logbook datasets. Nevertheless, there might

still be cases where the variables are not included and thus there is still a need to define these variables' states and then the conditional probabilities.

Discussion now moves on two issues. First is to do with approaches to estimate the parameters of the local distributions from the data. This is then followed with how to define the NPTs manually when there are no data and thus it is necessary to elicit the NPTs from SMEs.

4.3.3.1 NPTs Learnt from Data

There are two approaches of estimating the local parameters from data: the maximum likelihood estimates and the Bayesian estimates (Scutari and Denis, 2015a)

4.3.3.1.1 Maximum Likelihood Estimation

Through this approach it is possible to calculate the local, multinomial parameters from the empirical frequencies in the dataset using the BN structure (Nagarajan, Scutari and Lebre, 2013, chap.1).

However, the problem with this approach is when the dataset is sparse and there are state-cells in the tables that have zero counts. This is something that should be expected in the datasets dealt in the current research, since there are variables like those that describe certain components' failures to have very few counts of the failure level merely due to the components having been manufactured to be highly reliable.

4.3.3.1.2 Bayesian Estimation

A similar approach is to acquire the posterior distributions of the local parameters using a Bayesian approach. Applying uniform priors by assigning equal frequencies to the cells of the tables give:

$prior(x_i = k, pa(x_i) = j) = \frac{1}{r_i \times q_i}$, and $prior(pa(x_i) = j) = \frac{r_i}{r_i \times q_i}$

The corresponding observed counts in the dataset are:

$p(x_i = k, pa(x_i) = j) = \frac{number\ of\ observations\ in\ which\ both\ x_i = k\ and\ pa(x_i) = j}{n}$, and

$$p(pa(x_i) = j) = \frac{all\ the\ observations\ that\ pa(x_i)=j}{n}$$

In order to proceed, it is necessary to include an estimate about the importance/weight to include in the prior distribution and then get a Bayesian average. This is done by assigning a value to the *imaginary sample size (iss)* (also known as *equivalent sample size*) which expresses the weight wanted to assign to the prior as compared to the dataset when computing the posterior. The value of the imaginary sample size is divided by the number of cells in the NPT and then the posterior estimate is computed as the weighted mean of the (flat) prior and the empirical frequencies:

$$Bayes\ avg(x_i = k, pa(x_i) = j) = \frac{iss}{n+iss} prior(x_i = k, pa(x_i) = j) + \frac{n}{n+iss} p(x_i = k, pa(x_i) = j)$$

$$Bayes\ avg(pa(x_i) = j) = \frac{iss}{n+iss} prior(pa(x_i) = j) + \frac{n}{n+iss} p(pa(x_i) = j)$$

Finally, $p(x_i = k|pa(x_i) = j) = \frac{p(x_i=k,pa(x_i)=j)}{p(pa(x_i)=j)}$

The imaginary sample size is usually chosen small so that the prior distribution' weight can be easily dominated as the amount n of data increases (Scutari and Denis, 2015b).

4.3.3.2 Obtaining the NPTs from SMEs

When required to introduce values for the variables from SMEs, the principle of parsimony is very important to follow and especially for parent variables. Cain (2001) suggests aiming for two (preferably) or three values to describe the states of a variable and introduce more than three only in exceptional circumstances. The reason for introducing as few states as possible and especially for parent variables, is to keep to the minimum the number of conditional dependence probabilities needed to elicit from SMEs. If a child variable has two states and each of its N parents has k states then the number of parameters needed to elicit would be to the order of $O(k^N)$.

However, it is not only the number of NPTs that is a challenge that needs to be faced. The process of the elicitation of probability values can be quite challenging

too. The following observations, which have appeared in relevant case studies, summarise different problematic situations:

(a) Unless there is diversity among the nature of expertise, then the model might be biased towards the available knowledge (Garthwaite, Kadane, and O'Hagan, 2005).
(b) There can be cases in which SMEs want to support individual or group agendas (Garthwaite, Kadane and O'Hagan, 2005)
(c) Even though there is a good representation of experts, there might be still be a genuine disagreement on the estimated values (DeGroot, 1974; Douven, 2010; Garthwaite et al., 2005; Golub and Jackson, 2012; Hegselmann and Krause, 2002; O'Hagan, 1998)
(d) The time availability of the SMEs is limited (Linda C van der Gaag, Renooij, Witteman, Aleman, and Taal, 1999). In this very common problem, the time devoted to providing the needed numerical estimates depends on whether there is an interest / incentive by the experts to provide consultation. Furthermore, even if the interest exists, there is still the challenge of limited human attention span
(e) People estimate probabilities using certain heuristic principles. Even though these heuristics are very useful in reducing the complex task of assessing or predicting values and even though they are well related to the task, they can lead to systematic errors / biases (Garthwaite et al., 2005; Kahneman, Slovic, and Tversky, 1974). (see also Section 4.6.2 for a discussion on human biases and ways to deal with them)
(f) The level of the domain experts' familiarization to the concepts of probability, frequency as well as to specific measures like the mean, mode and variance is limited (Garthwaite et al., 2005; Gigerenzer, Hoffrage, Mellers, and McGraw, 1995; Wiegmann, 2005)
(g) If the number of variables is large, certain probability elicitation methods can be impractical (van der Gaag et al., 1999)
(h) There are variables on which there is very little experience/knowledge and therefore experts are reluctant on assessing frequencies (L C van der Gaag, Renooij, Witteveen, Aleman, and Taal, 2002)

In order to be able to deal with the challenges of obtaining parameter estimates from human experts, one needs to carefully choose the experts to be engaged and also the methods that are used for the elicitations. Furthermore, it is useful to be able to evaluate which of the nodes are more important for the model's use so that the focus is higher for their verification. This can be done by a sensitivity analysis study (Fenton and Neil, 2013, Chapter 8).

4.3.3.2.1 Choosing Experts

Regarding (a) above, Garthwaite, Kadane, and O'Hagan (2005) suggest that since the term "expert" emphasises the person who "*society and/or his or her peers attribute special knowledge about matters being elicited*", the combination of the expertise of several people might be able to relax the problem of individual agendas. Kaplan, Skogstad, and Girshick (1950) provide a very interesting and important observation: individuals who gave substantive reasons for their numerical forecasts had significantly better outputs than those whose reasons were either tautological or non-existent.

Apart from the obvious attributes of experience and studies, Cain (2001, p.32) and Clemen and Winkler (1999) provide some further guidelines to choosing experts:

- They must be accepted by the group they are representing
- They should possess good local knowledge
- The facilitator should be aware if there are possible financial or personal interests in the inferences or decisions (Garthwaite, Kadane and O'Hagan, 2005)
- They should differ from each other in terms of viewpoint and knowledge (Wiegmann, 2005)
- An optimum number is from three to five

4.3.3.2.2 Combining Expertise

When facing problem (b) in Section 4.3.3.2 of combining expertise, there is the dilemma of either letting the group interact openly and reach a consensus (behavioural aggregation) or consult each expert individually.

4.3.3.2.2.1 Working with Groups

A seminal work on the area in which the elicitation process is centred on using extensive SME interaction as its main mechanism, has been that of Budnitz et al. (1998). Core to their analysis is their categorisation in four types of consensus:

- Type 1 is when all experts get to have the same value for a variable or model parameter
- Type 2 is when all experts finally get to have the same probability distribution for a variable or model parameter
- Type 3 is when they agree that a certain composite probability distribution represents the specific group's expertise
- Type 4 is when the experts agree that a certain probability distribution represents the general scientific community

The authors' suggestion is to realistically seek either Type 3 or 4 and thus reach an agreement among the experts on *how to represent the spread and diversity* of their state of knowledge on a situation rather than *try to reach a consensus* of Type 1 or 2. Furthermore, they highlight that even though the effort is to represent the spread of the whole range of expertise, the engaged panels is just a sample. Consequently, care should be given to:

- A reasonable representation in the panel,
- Clear description of the technical basis for the assessments,
- A quantification of the uncertainties expressed,
- An effective peer review / validation of the outputs, and
- Documentation of all the above

Technically, the method aims to create conditions in which the different experts' outputs are equally weighted. In order to achieve the equality in weights, the process attempts to maximise the overlap through intensive and structured interaction among the experts. In order to achieve that, the authors suggest the following principles:

- The experts are viewed as informed evaluators of the expressed models rather than as supporters a certain (their own) position

- The interaction is not on named views but on hypotheses and underlying models
- There can be progress if the specific objects of disagreement are isolated
- The facilitator practices and encourages active listening by summarising and explaining back the points of agreement and disagreement.
- The facilitator clarifies that the purpose is not to find the correct answer or get to a consensus of the Type 1 or 2
- The facilitator states that the responsibility for the process of obtaining the aggregated probability distribution is the facilitators' and not the experts', but the intellectual responsibility lies to both. Therefore, the experts do not act only as evaluators but also facilitate the process to act as integrators as well

Furthermore, the authors provide a very interesting justification of why the occasional outlier expert whose interpretation is different from the rest but cannot support it with solid data or reasoning from the point of view of the rest of the experts, should be weighted low. Budnitz et al. (1998) suggest that the experts are a sample of the population of expertise and there can be a case that the sample is not representative.

4.3.3.2.2.2 Working with Individuals

The following discussion deals with the elicitation of probabilities and parameters without having the experts directly interacting with each other. Nevertheless, these methods can be used in combination with the group work (Budnitz et al., 1998).

Delphi method (Clemen and Winkler, 1999b; Pill, 1971) is a the widely known method that combines expert elicited values . The objective is to get values about something within the area of expertise of each and then combine these values to a final single that results from consensus. Anonymity, controlled feedback and statistical group response are its core features. Controlled feedback aims to help reach a consensus, while the statistical group response aims to let the participants understand that their opinions are represented in the final report. Furthermore, the ritual of a structured procedure is both a useful device for

facilitating the thinking and increases the level of acceptability of the results. The usual approach to combining the elicited values and to looking for a consensus has been to use either the groups' median or average and provide them back as a feedback hoping for reconsiderations and gradual revisions towards a consensus; since a single value per question is required, consensus is critical.

DeGroot (1974) suggested a variant of the method to help with the issue of verifying consensus. His approach involves asking each expert to provide both the requested value and a weight that reflects the importance of each of the participants including her/himself in the elicitation of the specific value. If there are k individuals asked, then create a $k\ by\ X\ by\ k$ matrix P where its rows are each individual and each column is the respective assigned weight. Providing $F^T = (F_1, F_2, \dots, F_k)$ values ($F^T$is the transposed vector of F) for the asked parameter, a combined output $F^{(1)} = PF$. Given mainly that the system of value revisions is assumed to form a Markov chain with P being its transition matrix, DeGroot (1974) showed that there is eventually a convergence/consensus as the number of revisions increase, $n \to \infty$. The resulting value of $F^{(n)} = P^n F$ is $lim_{n\to\infty} F_{i\,n} = F^*,\ i = 1,2, \dots, k$, i.e. each individual eventually ends-up with the same value F^*. The value is $F^* = \pi F,$ with $\pi = (\pi_1, \dots, \pi_k)$ the stationary probability vector whose components are calculated by solving the system of equations $\pi P = \pi$ and $\sum_{i=1}^{k} \pi_i = 1$. The assumptions suggested by DeGroot (1974) for a consensus to be reached were:

- There is at least one column with positive integers in at least one of the P^n iterations. This assumption says that there is at least one individual whom all participants recognize that her opinion has a non-zero weight
- During the iterations, the individuals are informed of the values assigned by others and as a result change their initial values F_i for the parameter asked in such a way that the revised value is a linear combination of everybody's values and her assigned weights. So, after the first iteration, individual's i revised value would be: $F_{i1} = \sum_{j=1}^{k} p_{ij} F_j$, with p_{ij} the weight that he/she has assigned to individual j

- In the iterations that follow, each individual finds no reason to be inconsistent with her original weights assignment distribution

Other studies relaxed the assumptions above, and resulted in non-linear problems which, owing to their complexity, they investigate the process using simulations (Douven, 2010; Hegselmann and Krause, 2002). However, due to the time constraints of the SMEs and the convenience provided by the linear assumption, the results of Clemen and Winkler (1999b) are more practical to apply. Furthermore, DeMarzo, Vayanos, and Zwiebel (2003) and Chandrasekhar, Larreguy, and Xandri (2015) show that updating behaviour is better captured by the repeated averaging models than by more complicated rules. The assumptions needed in order to increase the probability of convergence are that the number of individuals is large and there is no agent that is too popular or influential.

When coming to the problem of how to elicit the numerical values from each SME, there are many issues regarding the way that people think and feel. Individuals estimate the likelihood of an event by the use of certain heuristics. The problem though is that these same heuristics which help individuals provide the information are prone to biases like the overconfidence bias, insensitivity to sample size and to prior probabilities, misconceptions of chance etc. (Fenton and Neil, 2013). There are a number of methods designed to supress the effect of biases (Wiegmann, 2005):

- Gamble methods: The individual is presented with a lottery on which the outcome depends on the probability set by the facilitator and a lottery on which the outcome depends on the probability of what is asked. The individual is then asked to choose. The facilitator changes the probability of the first lottery's outcome until the individual asked is indifferent between either of the two choices. The resulting probability is the finally elicited value. The problems that this approach presents are that it can be time-consuming, depends on the risk attitude of the individual, and can be sometimes difficult to conceive especially when the asked questions are about very rare events. The risk attitude effects can be reduced if instead

of two lotteries, the individual is presented to a spinning wheel - a pie chart - with the percentage of the areas of each event equal to the tested probability. However, this method is still prone to certain heuristic biases and it is again time consuming.

- Hierarchical methods: Using the Analytical Hierarchy Process (AHP) (Monti and Carenini, 2000; Saaty, 1980) the probabilities are elicited by comparing the chances of each possible pair of events. AHP method in general provides a measurement of the consistency of the results and also the individual is given that measurement to refine her outputs. However, there are many pairs that need to be compared and some of them are too different and thus hard to analogise. In the same class of methods, Druzdzel and Gaag (2000) developed an approach which recognizes that the information might exist in either a qualitative or a quantitative form and with different magnitudes of precision. The basic idea is that the true distribution of probability values of a variable lies within a distribution hyperspace of all possible probability distributions. The individual can provide either a qualitative or a quantitative estimate depending on which of the two would make her/him more comfortable. Information provided expresses a constraint and under the assumption that all probability distributions that are compatible with the provided information are equally likely, they define a system of (in)equalities and derive second-order probability distributions to determine the most likely one.
- Frequency estimation methods: Various studies have indicated that individuals find it easier to provide values estimated in a frequency format rather than in a probability format (Gigerenzer et al., 1995). The individuals are asked to provide the number of times that they would expect an event to occur out of some multiple of ten, usually on a graphic scale that is fast and easy to understand. Linda C van der Gaag et al. (1999) realised the need for either a quantitative or qualitative alternative to express the values, and developed a scale that uses both numerical and verbal anchors on each side. That scale was used to elicit the probabilistic information sought by the experts and provided the additional benefits of

being easy to use, closer to their usual cognitive processes and being able to handle multiple variables in a relative small time. However, this scaling approach occasionally tends to be inaccurate and prone to scaling biases.

In the problems of interest to the present research, the experts' availability and the easiness of use of the method are very important factors in choosing an eliciting approach. An important pre-step to increasing the reliability of the outputs is to train the experts before the actual application of the method (van der Gaag et al., 2002). This can be accomplished by explaining and trying out the approach on familiar, everyday topics with the objective to verify that the approach is equally well understood by all participants. Another important consideration is the way that the scenario's context with the values of the parents are presented to the experts for their estimate on the (conditional) frequency of the value of a child. The fragments of texts used to describe what is required by the experts, are easier to relate to when presented as likelihood / chance instead of frequency. At the same time, it is helpful for the comparison to present the whole NPT values needed on the same page.

Of course, any subjective method has the risk of being inaccurate. However, the level of inaccuracy depends on the problem's context and the problem itself, so a sensitivity analysis would reveal the variables that are of interest to focus the data collection efforts more (Fenton and Neil, 2013, Chapter 8). The suggested approach then is to use the elicitation of the values using the probability scale with numerical and verbal anchors as a first step, execute a sensitivity analysis and then refine the results of the variables that are of higher interest.

4.3.4 Validating the BN model

The BN model, after being populated by the NPTs needs to be verified and validated to see if the BN is a faithful model, i.e. whether the BN graph represents the independencies that exist among the variables and also that it represents the dependencies that exist. Approaches to establish such an evaluation include (Mahoney and Laskey, 1996; Neil et al., 2000):

- See if key variables' marginal NPTs match known distributions
- Compare opinions from different SMEs

- Overall review of the definitions of the nodes, their states and the independence assumptions
- Do an importance analysis of a number of selected focal nodes in order to identify the effect of changes of the values of / to peripheral nodes
- Do the above using different scenarios
- Do the above using typical, infrequent, and unanticipated conditions

4.4 Discretisation

4.4.1 Variables in the BN Models

The variables that can potentially participate in the demand forecast models are not just categorical like the environmental conditions or the types of inventory policies, but can also be also numeric. These include continuous variables, with some of the obvious ones being the operating hours that a system's component has, the duration of the repair activities and the duration of the resupplying of the inventory. Furthermore, in the non-categorical set of variables, one can also have discrete variables, like the number of repair equipment that are installed in a depot and the number of mechanics that are working.

There are a number of advantages in keeping the numerical nature of the variables. One important advantage is that the numerical variables carry more information than the categorical. The numerical variables have the attributes of the categorical, i.e. each value is a different "state", but additionally each value has a certain order as compared to any other and finally, two different values have a different distance from a third. However, a number of the common applications of BNs require that all variables used are categorical, i.e. they have a finite set of discrete states.

There are of course applications of the BNs in which the variables are indeed numeric. One such case is the Gaussian Bayesian Networks. In these networks all the variables are assumed to follow the normal distribution while the network expresses a Multivariate Gaussian Distribution. Nevertheless, Gaussian Bayesian Networks make the assumption that all modelled variables are continuous and follow a Normal distribution and thus cannot be used for the

present research cases where the set of factors can also include categorical variables or cannot be well approximated by the normal distribution (e.g. the time to experience a failure for a component). On the other hand, there are the Hybrid Bayesian Networks that can handle a mixture of variables.

Hybrid[12] Bayesian Networks are a very powerful type of BN models due to their ability to handle a mixture of types of variables and thus extend their applicability (Margaritis, 2003; Stefano Monti and Cooper, 1998a; Scutari and Denis, 2015a, Chapter 2). The joint probability distribution that such a model expresses can be estimated by the use of Markov Chain Monte Carlo (MCMC) methods (Fenton and Neil, 2013, Chapter 9; Scutari and Denis, 2015a, Chapter 3). Another equally interesting approach to combining the mixture of variables in a Hybrid BN is by the use of Dynamic Discretisation (Fenton and Neil, 2013, Chapter 9; Neil, Tailor, and Marquez, 2007). However, the learning of the Hybrid Bayesian Networks structure from data is not as efficient as the equivalent of the conventional BNs in which all variables are multinomial (Stefano Monti and Cooper, 1998b), and it has not been widely used (see Monti and Cooper (1998a)). Consequently, there is still a need to rely on subject matter expertise for the development of the structure. Nevertheless, as shown earlier in Chapters 2 and 3, it is not easy to find SMEs with a solid understanding of the width and spread of a multifunctional system such as the SC.

What now follows, is a discussion about modelling challenges in the use of numeric and categorical variables in models like BNs. This is followed by a discussion about the different methods of discretisation that were considered for the problems being investigated in this research.

4.4.2 Discretisation Challenges

A general ascertainment is that numeric variables' investigation can be problematic due to the number of degrees of freedom that can inherently exist in

[12] Not to be confused with the use of the term "hybrid" in some of the models that developed in the present research (for the list of models see Sections 1.4, 7.2.3 and 7.3.3). In those cases the term refers to the BN's DAG structure-identification approach, while here it refers to the type of variables used

arbitrary numeric distributions (Hartemink, 2001). This fact increases the dimensionality of the problem of learning the structure of a BN model from data and thus is in the root causes of the increased difficulty in doing so when numeric variables are included in the variables' set. Furthermore, given that the amount of usable data that might be available can potentially be comparatively limited, there is a need to find ways of reducing the dimensionality of the models that are developed.

The dimensionality can be reduced either by making parametric assumptions about the distribution of the numeric variables, or by discretising them into a small number of intervals where each interval is mapped to a level/category. However, choosing the discretisation alternative means that the information inherently carried by the numeric variable, namely the ranking among the values and the relative distance between them, is lost since the mapping produces simple categories with the only attribute maintained being that they are mutually exclusive. Furthermore, the resulting model might not be sensitive to new situations that have not been included to the dataset that has been used for its development. Nevertheless, the preference here is to use discretisation in the development of the BNs for the following reasons (Hartemink, 2001)`:

1. Discretisation reduces the dimensionality of the problem and thus unsupervised learning algorithms can be used in order to provide a mapping of (hopefully) the majority of the relationships among the variables to be modelled
2. The records kept in the logbooks are just snapshots of the variables' values when that recording was made. Eventually, in order to preserve the continuity of the numeric variables there is either a requirement to make distributional assumptions or if empirical distributions are preferred, then there is a need to interpolate between the recorded values or to extrapolate beyond their range
3. Discretisation in general introduces a means of robustness against error that can arise during measurement or recording
4. It seems that the relationships between some of the numeric variables and related categorical ones can stimulate reasonable approximations of their

values by the use of qualitative statements. So, for example certain ranges of hours worked for a component can be characterised as “High” as compared to the related variable that includes the event of a breakdown and to the variable of preventive maintenance that includes the event of a component’s replacement. Certainly, such characterisations can be considered as simplifications and that other non-linear relationships over the numeric spectrum would be able to distinguish more detailed interactions between the factors. However, a good discretisation is likely to capture most of the qualitative sense of the relationships and indeed this is what the discretisation algorithms try to do

5. Discretisation can be the initial step that can help us understand the relationships and build a first model, but then it can be followed by another iteration that uses more intervals or even consider numeric levels
6. When it is not clear which distributions to choose in order to model the numeric sampled variables then discretisation has the potential to produce better models

4.4.3 Definitions

While wanting to limit the range of the numeric variable to a set of discrete categories, these categories need to have a (discrete) numerical mapping in the set of integers. The reason being that it is desirable for the categories to have numerical values related to them in order to develop and use algorithms that optimise the number and the boundaries of these categories.

Therefore, the following definition (Hartemink, 2001) is provided in order to be used in the discussion of the methods: If $x \epsilon \mathbb{R}^N$ (or $x \epsilon \mathbb{Z}^N$) is a sorted, real valued vector with size N, its *discretisation* is an integer vector d that has the same length as x and which satisfies the following properties:

- For some integer $D \in \mathbb{N}^*$ each element of d is in the set of $\{0, \dots, D-1\}$
- $d_i \leq d_j$ if and only if $x_i \leq x_j$ for all i, j

If additionally the first element in the space of D is equal to 0 and the last is equal to $D-1$ then the discretisation is also called *spanning* (Hartemink, 2001).

The *discretisation policy of degree D* is defined as a vector Λ with real values that has length $D + 1$ and has the following properties:

- Its elements are all ordered, i.e. $\forall\ i < j, \Lambda_i < \Lambda_j$
- It includes all the real numbers, i.e. $\Lambda_0 = -\infty, \Lambda_D = +\infty$

In essence the elements of Λ delineate the left and right boundaries of the D intervals of the discretisation. Consequently, a discretisation under the defined policy results in the following mapping of the real vector x to the integer vector d:

$$\Lambda_j < x_i \leq \Lambda_{j+1} \Leftrightarrow d_i = j, \qquad \forall\ i \in \{0, \dots, N-1\}, j \in \{0, \dots, D-1\}$$

4.4.4 Interval Discretisation

*Interval discretisation (*or *equal-width intervals)* is a simple type of discretising numeric variables. In the interval discretisation, the area between the 1st and last element of x, i.e. $[x_{0,} x_{N-1}]$ is divided in D equal intervals, irrespective of the in between values, and then the elements of x are distributed between these intervals:

$x_0 + \frac{j(x_{N-1}, x_0)}{D} < x_i \leq x_0 + \frac{(j+1)(x_{N-1}, x_0)}{D}$, for $i \in \{0, \dots, N-1\}$ and $j \in \{0, \dots, D-1\}$ respectively.

The boundaries of the above discretisation can be expressed as following policy vector:

$$\Lambda = (-\infty, x_0 + \frac{x_{N-1} - x_0}{D}, x_0 + 2\frac{x_{N-1} - x_0}{D}, \dots, x_0 + (D-1)\frac{x_{N-1} - x_0}{D}, +\infty)$$

4.4.5 Quantile Discretisation

Quantile discretisation (or equal-frequency) is another simple type of discretising numeric variables. In contrast to the interval discretisation, in this method the number of values between the 1st and last element of x are taken into consideration and thus D intervals are chosen in such a way that there are equal number of observations in each. In essence the choice of the interval is defined by the index of the x's element given of course that the elements are sorted. So, if the real vector x is of size N, the i^{th} observation is allocated to interval j^{th} as follows:

$\left\lfloor \frac{jN}{D} \right\rfloor < i \leq \left\lfloor \frac{(j+1)N}{D} \right\rfloor$, for $i \in \{0, \dots, N-1\}$ and $j \in \{0, \dots, D-1\}$ respectively.

The boundaries of the above discretisation can be expressed as following policy vector (observe that the boundaries are now defined by the indices of x through the ratio of N the number of elements in x over D the degree of discretisation):

$$\Lambda = (-\infty, \frac{x_{\lfloor \frac{N}{D} \rfloor} + x_{\lfloor \frac{N}{D} \rfloor + 1}}{2}, \frac{x_{\lfloor \frac{2N}{D} \rfloor} + x_{\lfloor \frac{2N}{D} \rfloor + 1}}{2}, \dots, \frac{x_{\lfloor (D-1)\frac{N}{D} \rfloor} + x_{\lfloor (D-1)\frac{N}{D} \rfloor + 1}}{2}, +\infty)$$

Contrasting these two simple approaches, quantile discretisation takes into consideration only the order of $x's$ elements while interval discretisation accounts for their distance as well. However, interval discretisation creates intervals of equal length. On the other hand, this same attribute can lead to certain allocated value areas not to be represented at all by the data in x.

These algorithms can produce a reasonable abstraction of the recorded numeric data (Kerber, 1992). However, in multivariate studies the variables are often not considered individually. For example, the numeric variable *number of hours that a subsystem or component has operated* is related to the categorical variable *preventive maintenance incident* since the former can be used to inform preventive maintenance activities. If the modeller decides to discretise the numeric variable without considering its context-relation to the categorical, then some of the mutual information will be lost. The same challenge is faced when more two or more numeric variables that are associated to each other. An example of such a case is between the numeric variables *duration of a resupply order* and *duration that a component stays in the repair shop*. In this case, the latter is affected by the former and thus, discretisation should take their relationship into consideration.

Consequently, doing discretisation in isolation does not preserve the predictability of the one variable over the other, which is information that existed before discretising them. There are two general approaches to discretising numeric variables by considering their associations with other variables; static and dynamic.

Static discretisation is performed as a preparation step before the development of the BN. The main advantages of a static discretisation are that the work needs to be done only once and that unsupervised BN-structure learning algorithms can be applied with fewer computational limitations (Hartemink, 2001; Stefano Monti and Cooper, 1998b) since they have to deal with only the unchanging multinomial distributions for all variables involved. On the other hand, with the dynamic discretisation the benefit is that it is not as restrictive as the multinomial distributional assumption for all the variables and therefore the BN model can better preserve the information carried by them (Aven, 2016), given of course that the distributional assumptions for all the variables are correct. However, as mentioned earlier, the existing unsupervised BN-structure learning algorithms that use datasets which are both numeric and categorical are not as efficient (Hartemink, 2001; Stefano Monti and Cooper, 1998b). Moreover, to the best of the author's knowledge, most of the commercially available BN structure unsupervised learning packages require all the variables to be categorical/discretised in a static manner.

The following sections discuss a number of algorithms which perform discretisation of numeric variables by associating them to a categorical (*ChiMerge* and *mdlp*), and of numeric variables by associating one of them to the rest (*Hartemink*), with the aim to preserve the relative information carried in each.

4.4.6 ChiMerge Discretisation

ChiMerge algorithm (Kerber, 1992) is one of the first algorithms that performs the discretisation as a pre-processing step (i.e. statically) and not dynamically as the model-development algorithm runs.

The idea is that a sorted vector $x \epsilon \mathbb{R}^N$ (or $x \epsilon \mathbb{Z}^N$) that needs to discretise by associating it with an equally sized set C that includes elements of categorical nature. The number of categories included in C is c. The vector x is initially discretised by placing one interval for each of its N elements (i.e. initially $N = D$). From then on, the algorithm is composed of two steps repeated sequentially:

1. For each adjacent interval the χ^2 value is computed as follows:

$$\chi^2 = \sum_{k=1}^{2} \sum_{m=1}^{c} \frac{(O_{km} - E_{km})^2}{E_{km}}$$

Where:

k is the index of the number of intervals that are compared (the adjacent 2)

m is the index of the number of classes

c is the number of classes of the categorical variable that is associated to x

O_{km} is the number of x elements in k^{th} interval that correspond to the m^{th} class

R_k is the total number of x elements in k^{th} interval, i.e. $R_k = \sum_{m=1}^{c} O_{km}$

L_m is the number of x elements of the m^{th} class in both intervals, i.e. $L_m = \sum_{k=1}^{2} O_{km}$

n is the total number of x elements in both intervals and it is equal to $n = \sum_{m=1}^{c} L_m$

E_{km} is the expected frequency of the x elements in k^{th} interval that correspond to the m^{th} class, i.e. $E_{km} = \frac{R_k L_m}{n}$

2. After calculating all χ^2 values, merge the pair of adjacent intervals with the lowest χ^2

The algorithm stops when a user-defined threshold χ^2 value has been exceeded.

In this algorithm - it can be observed that, the use of the χ^2 test is partly driven by the chosen bottom-up approach, i.e. to start by creating a partition for all the elements of x. There are several static discretisation algorithms that have been developed (Chmielewski and Grzymala-Busse, 1996; Gonzalez-Abril, Cuberos, Velasco, and Ortega, 2009; Huan Liu and Setiono, 1997). The algorithm describe next follows the reverse approach which is top-down. It is an efficient algorithm that has is applied in the *R* environment through the *discretise* package (Kim, 2012), using the *mdlp* (Minimum Description Length Principle) function.

4.4.7 MDLP Discretisation

Once again, it is assumed that a sorted vector $x \epsilon \mathbb{R}^N$ (or $x \epsilon \mathbb{Z}^N$) and a related set C with categorical values exist. The *mdlp* algorithm chooses a cut-point T that

divides x into two parts by looking for a candidate T through the evaluation of the whole range of points in x. The candidate cut-point T is placed between every pair of values in x and therefore there are $N-1$ evaluations in order to pick the "best". The criterion that is used in this algorithm is the *information entropy* that the algorithm tries to minimize in x given C.

If $p(C_m, S_1)$ is the proportion of elements in the subset $S_1 \in x$ that are associated to the class C_m, then the entropy of this subset S_1 is defined[13] as $H(S_1) = -\sum_{m=1}^{c} p(C_m, S_1)\log_2 (p(C_m, S_1))$. What the specific metric provides can be understood if the concept of "entropy" is taken as the "lack of predictability" (Oxford University online dictionary, 2018b). Entropy, in the present research' context of associating a class C_m to a specific value in S_1, can be understood the "surprise" of seeing such an association. Consequently, the higher the estimated probability (proportion) $p(C_m, S_1)$ the lower the surprise. Therefore, since the higher $-\log (p(C_m, S_1)$ is, the higher the "surprise", the measure of the entropy $H(S_1)$ expresses the average of the "surprise" of having a certain set S_1.

Accordingly, the algorithm searches for the specific T that partitions x in S_1 and S_2, so that their average entropy is minimised: $\min(H(x,T) = \frac{|S_1|}{|x|}H(S_1) + \frac{|S_2|}{|x|}H(S_2))$.

One could think that the algorithm is inefficient since it searches for T $N-1$ times as it was stated earlier. However, as Fayyad and Irani (1993) show, a value for T found through the minimisation of $H(x,T)$ is always between two different classes and thus the number of searches are reduced since it leaves out searches in subareas of x that are associated to the same class.

The top-down approach continues by bisecting the parts through recursively applying the algorithm to the optimum subsets S_1 and S_2. In contrast to the *ChiMerge* that were discussed earlier, the *mdlp* algorithm uses the Minimum Description Length Principle criterion to stop the partitioning (Fayyad and Irani, 1993; Liu, Hussain, Tan, and Dash, 2002; Ross Quinlan and Rivest, 1989). The

[13] The same notation is used as in the previous algorithm

criterion suggests that partitioning of x does not continue when the information gain as estimated by $G(x,T) = H(x) - H(x,T)$ - i.e. the algebraic difference between non-partitioning x and partitioning it in S_1 and S_2 – is less or equal to $\frac{log_2(N-1)}{N} + \frac{log_2(3^c-2)-[cH(x)-c_1H(S_1)-c_2H(S_2)]}{N}$, with c the total number of classes corresponding to the whole x, c_1 and c_2 the number of classes corresponding to S_1 and S_2 respectively.

The previous two algorithms that were partitioning the numeric vector x by using its association to a categorical vector C. However, in the problem cases that are examined in the present research, there are variables that are associated and which can be only numeric. For example, the hours flown of two components of a system are both numeric and are associated since they both refer to the usage of the same system. In such cases what is wanted is to preserve as much of the mutual information that exists between the pair of variables as possible. For these cases the present research has adopted *Hartemink's* discretisation algorithm (Hartemink, 2001) that is also embedded in the *bnlearn* package in R (Nagarajan et al., 2013; Scutari and Denis, 2015a), one of the packages that was used in the present study in order to build the BNs.

4.4.8 Hartemink Discretisation

Similarly, to algorithms like *ChiMerge*, *Hartemink's* algorithm uses a bottom-up approach and it is also a supervised algorithm in the sense that the user needs to choose and define certain attributes. For reasons to be subsequently explained, n this specific case the attributes that the user needs to choose are the stopping criteria and the number of initial intervals.

In order to be able to deal with the set of variables under consideration that are numeric, the user needs to define a number of intervals/breaks that each of the variables is going to be partitioned initially using quantile discretisation. The choice of quantile as compared to interval discretisation is supported by the fact that quantile retains more information (Hartemink, 2001). The number of breaks is usually chosen to be large so that there is not any unwanted merging between the elements of the variable before the algorithm starts (Scutari and Denis, 2015a).

Subsequently, the algorithm is composed of two loops. There is an outer loop which just reduces the number of intervals until the final number which is predetermined by the user. This has the role of the stopping criterion. The inner loop is the one in which suggested merging is evaluated and eventually decided using the Total Mutual Information score as a metric for the evaluation.

The Total Mutual Information score is defined as the sum of the mutual information between all pairs of variables (Hartemink, 2001). The mutual information among two variables X and Y is a metric that is defined as:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

As compared to the entropy mentioned earlier, the mutual information metric instead of the logarithm of the joint probability, uses the log of the ratio of the joint over the two marginal. If the two variables are independent this ratio is equal to unity and therefore the logarithm is zero, an output which expresses the lack of mutual information due to the independence. Consequently, the ratio compares the joint probabilistic relation of the two variables over their joint relation under the assumption of independence.

At each step with the inner loop, for each variable in the set, each pair of the neighbouring intervals are merged in turn and the Total Mutual Information score between the variable under consideration and all the rest of the variables that are to be discretised is calculated and it is compared to the respective value if the merging would not take place. The pair that is chosen to become a single interval in the specific variable is the one that produces the lowest reduction to the value of the Total Mutual Information. In order to avoid the effect of the order that the variables are worked, none of the merging takes place before all the variables have been considered.

4.5 Logistic Regression

4.5.1 Characteristics of the (Binary) Logistic Regression Models

The (binary[14]) logistic regression model belongs to the family of multiple regression models but with a response variable that is categorical and can take only two values/categories, as compared to the multinomial or polychotomous logistic regression with which one wishes to forecast the membership of more than two categories.

In a general multiple regression model, the response variable Y_i can be estimated by the straight line formed from the sum of a given set of covariates X_{ji}:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \varepsilon_i$$

, in which the b_j is the regression coefficient of the corresponding covariate X_{ji}. In logistic regression, instead of providing a forecast for the value of the response variable given the values of the model's covariates, the forecast is for the probability of occurrence of one of the two categories of Y. In essence, the model gives the exponent in the following function:

$$p(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni})}}$$

In order to be able to use the linear model of the common regression model, instead of the binary-categorical output, use the natural logarithm of the ratio of the odds of experiencing one of the two values over the other. The name of this logarithmic ratio is $logit$:

$$logit(p) = \log\left(\frac{p(Y)}{1 - p(Y)}\right) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni}$$

Eventually, given that the output of the above model is the $logit$, take its exponent on the natural base to get the odds ratio, which can then give the $p(Y)$

[14] For the rest of the present work, logistic regression refers to the binary

4.5.2 Assumptions

The logistic regression has similar assumptions to the general multiple regression:

- Linearity: assume that the $logit$ can be modelled by the linear relationship of its predictors
- Independence of errors: This is a general assumption for the datasets that is required to be used in the BN models built for the current research as well. It suggests that the cases within the dataset are not related (are exchangeable - for a discussion on exchangeability see e.g. Gelman et al (2014, Chapter 1))

4.6 SMEs' Judgemental Adjustments of a Model's Forecasts

The adjustments of a model's forecasts by the SMEs is something considered in the present thesis due to its wide applicability in industry (Christopher, 2016; Fildes et al., 2009; Makridakis, Wheelwright, and Hyndman, 2008, Chapter 11), and also due to the fact that decision makers tend to adjust, given knowledge about the future context (Fildes et al., 2009).

4.6.1 Single Exponential Smoothing (SES)

The forecasting models that are commonly used in the industry are types of time-series statistical models. There are several reasons that this practice is common. Firstly, time series are relatively easy to produce since they usually rely on simple functions. Secondly, there is no need to store a large amount of data in order to train models and update them. Thirdly, given the very large number of different spare parts for which a forecast is needed, a common practice is to forecast in segments according to their common historical demand behaviour. This is something that is done very easily when only past time series demand data are used as predictors of future demand.

In the cases examined in the present research, it was decided to use the Single Exponential Smoothing Model (SES). SES is a very commonly used time-series model which has occasionally demonstrated very good results in forecasting competitions (Makridakis and Hibon, 2000).

$$\bar{d}_{t+1} = \alpha d_t + (1-\alpha)\bar{d}_t$$

, where $\bar{d}_{t+1}$ is the forecast demand value provided for time $t+1$, d_t is the demand experienced at time t and α is the weight considered for the experienced data point. The current research uses the per month recorded number of demands of the training dataset in the “tsintermittent” package in R.

4.6.2 Forecasts’ Judgemental Adjustments

Given the amount of changes that can take place during the final phase of the operations, one would expect that at least the uninformed forecasts provided by SES to be judgmentally adjusted by SMEs.

As Makridakis, Wheelwright, and Hyndman (2008, Chapter 10) state in judgmental adjustments the challenge is to combine the statistical outputs with the best aspects of the SMEs’ judgements, while at the same time avoiding the human biases.

4.6.2.1 Dealing with Judgmental Biases

There are a number of judgmental biases that can affect judgemental forecasting (Goodwin and Wright, 2014, Chapter 10; Makridakis et al., 2008, Chapter 10).

The first bias is that of "inconsistency”. This bias refers to SMEs changing their decisions when there is no clear reason, or in other words, they are unwilling or unable to apply the same criteria or procedures when making similar decisions. The reasons can be multiple: not being able to recall the criteria that they used, or the steps taken in the past, being influenced by mood, wanting to try something new or even explaining some signals as indicators of an influential change in the forecast’s context when in reality there was no such case.

The bias of “inconsistency” can be reduced by formalising the processes used for the decision making. This can be done by identifying first the influential factors, give them a weight of importance to the forecast and evaluate them. The latter step is indeed very important. It is the monitoring that can compare historical performance of the rules to the new situation and thus identify possible trends. Furthermore, the evaluation of whether a calibration worked, or it did not should be applied in order to activate and maintain the process of learning. Indeed, if learning does not take place, there is a different bias, that of “conservatism”.

“Illusory correlations” bias is when the SME considers the associations between the output of interest and some factors, when they are really just correlated through a confounding factor. This bias can also be accompanied by the “search for supportive evidence” bias in which the SME chooses which facts support a certain output while disregards the rest. A way to reduce them is similar to what it was discussed above, i.e. by verifying the patterns of the relations between the variables that are thought to be influential and the forecast output.

“Optimism” is another bias in which people prefer and thus forecast what they think is better for them, and which can also lead to underestimating the uncertainty about the future (“underestimating” bias). Diversifying and increasing the number of participants in the forecasting process can help in the reduction of such a biases.

Combining a judgements approach to reducing biases can be done either with each SME in isolation (see earlier discussion on the Delphi method in section 4.3.3.2) or as a group. The first approach is too time-consuming for the required purposes of getting demand forecasts for many components. On the other hand, the second can introduce a different kind of bias, that of groupthink. The “groupthink” bias appears when the members of a group tend to be supportive of dominant personalities and of each other, in order to avoid conflict during the meetings (Makridakis, Wheelwright and Hyndman, 2008).

A related bias to group thinking is that of "success/failure”. This takes place when one believes that either of the two is attributable to unique personal qualities. Encouraging the benefits of learning from errors is a way that can help towards to reduction of such a bias.

“Recency” bias appears when more recent events are given a higher weight than the older events, while another related is that of “availability” in which the events that can more easily be recalled are given a higher weight. Both of these biases can be reduced if a sound argument is presented to support the suggestions.

This remedy approach can be applied to the "anchoring” bias as well in which the SME is influenced by improper initial information. Furthermore, in order to reduce

the anchoring bias, SMEs can be presented by being initially given the model's forecast.

"Regression effects" can also appear when just by chance there can be persistent unidirectional changes which can be considered as indications of an existing trend. In such cases, one could try to support both the "is" and "is not" arguments in order to see whether to support the case of a true change.

Finally, another important judgemental bias is that of "conventional Wisdom" when this is supported just by unfounded beliefs. Again, developing an argumentative causal chain that is verifiable can help in the challenge of such a bias.

4.6.2.2 Combining Model Forecasts with Judgements

Makridakis et al. (2008, Chapter 10) suggest starting with reducing the anchoring bias by giving each of a group of SME participants a folder with values relevant to the forecast. The participants are made clear that even though this is historical information, the future might not be the same. The pre-work continues by asking the participants to write down the factors that they think can affect the output of interest and then adjust the statistical forecast anonymously. The factors and the thinking that supports them can be used as a formalising process and learning tool.

In the meeting that follows the participants agree on the value they want to acquire. In the cases used in this thesis the average of the values provided by the experts was used (Fildes et al., 2009).

4.7 Conclusions

Chapter 4 presented the BNs' DAG development methods that have been examined in this thesis, as well as the methods for estimating their NPTs. It was argued that due to the nature of the data in the FPPs, score-based algorithms are preferable for the DAGs that are developed through machine learning, while for the same reason Bayesian estimation is suggested for the NPTs. Additionally, methods for eliciting the DAG from SMEs using a number of idioms were also presented.

Furthermore, the BNs employed assume multinomial variables, and therefore methods for discretisation of continuous variables have also been examined.

Finally, the other two forecast methods used for comparison in the current thesis have been presented, namely the logistic regression and the SME's judgmentally adjusted forecasts.

5 ACCURACY AND ACCURACY IMPLICATION METRICS

5.1 Introduction

This chapter is divided in two parts. Firstly, a number of accuracy metrics are reviewed and evaluated for their applicability in the FPP. There are two reasons for which it is required to be able to evaluate and choose appropriate accuracy metrics. The first reason is that the datasets that were used were outputs of the demand for not just a single type of spare part. Therefore, one would expect different magnitudes of demand and to be able to use accuracy metrics that can accommodate such a requirement. The second reason is that the datasets were from the multiple runs of every simulated future scenario (see Section 1.3), which means that they were different sets of time-series, and this was an additional challenge to comparison of the models' outputs. Consequently, for the above two reasons, and in order to develop a better understanding of the accuracy metrics' possible limitations, an algebraic analysis of them was performed.

In the second part of the current Chapter, some accuracy implication metrics are reviewed and evaluated. As discussed then, a decision maker is mostly interested in how well a demand forecasting model can help with inventory related decisions, and the evaluation of how well each model contributes to such decisions is made by the accuracy implication metrics. Given that the FPP is a rather new type of problem (see Section 1.2), in the second part of this Chapter a study is performed on the required accuracy implication metrics. As is shown for the FPP, the existing metrics need to be complemented by additional ones.

5.2 Evaluating the Forecast Models

Evaluating alternative forecasting models is a challenging task and this can be inferred by the number of different accuracy metrics that exist in the literature and by the still existing lack of an omni-acceptable metric especially for demand data with intermittent behaviour (Kourentzes, 2013). Nevertheless, one of the core questions that need to be addressed when in the process of choosing the ways to compare and evaluate forecast models is whether forecast accuracy is an end in itself or is it a means towards an end (Boylan and Syntetos, 2006). The answer

depends on the stakeholders' objectives. The stakeholder can either be a model developer whose objective is to compare forecast models for their performance in the context of accuracy competitions (Makridakis and Hibon, 2000), or wants to down-select among a number of models that are then evaluated. In such cases, the accuracy of the forecast model is an end on its own. However, the forecasts are also outputs that are used as inputs to inform decisions, and thus, should be evaluated by the relative level of value that they bring to their areas of application (Gelman et al. 2014, p.142).

The applications of interest to the current research, are on forecasting the demand for spares. The forecast models in such cases aim to produce outputs that are accurate enough towards lowering the different stock holding costs and at the same time improving the level of service provided by the holding of spares. These are two objectives which are often competing. The stock of spares is held in order to contribute to the increase of the availability of the systems that need them when they are maintained/repaired, by reducing the logistics delay time, i.e. the factor $MLADT$ on the right of the denominator in the following function for the Operational Availability A_o metric:

$$A_o = \frac{MTBM}{MTBM + MTTR + MLADT}$$

$MTBM$: Mean Time Between Maintenance activities (either corrective or preventive)

$MTTR$: Mean Time To Repair

$MLADT$: Mean Logistics and Administrative Delay Time

On the other hand, the competing objective calls for having minimum costs when holding stock. These costs include the duration dependent costs of holding inventory, the stock obsolescence costs, the re-order costs, the unit purchase costs, the costs of backordering and the administrative costs (Axsater, 2006, Chapter 3; Hadley and Whitin, 1963, Chapter 1).

Consequently, accurate forecasts can contribute to the above two objectives, but the final improvement is also driven by the inventory rules that are applied, i.e. how the forecasts are used in order to determine "when" to place a stock

replenishment order and "how much" to order at each replenishment (Boylan and Syntetos, 2006; Syntetos, Nikolopoulos and Boylan, 2010). This interaction between inventory rules and forecast accuracy performance can be observed in the following Figure 5-1 (Boylan and Syntetos, 2006):

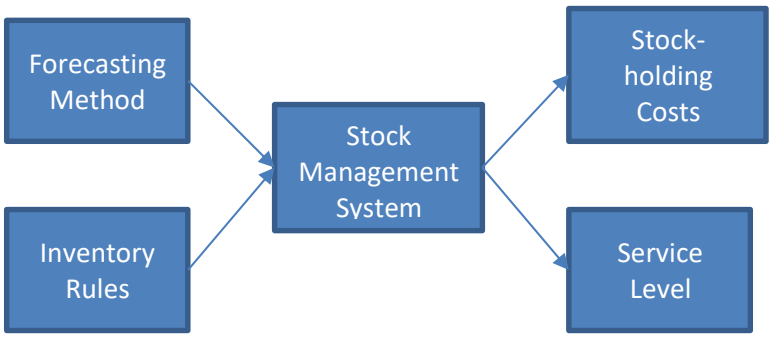


Figure 5-1: Effects of the forecasting methods and the inventory rules on the Service Level and Stock-holding Costs

As mentioned earlier, in order for the modeller to decide which model to use, he/she needs to assess the accuracy of one model as compared to others and therefore would need metrics to do this directly.

The paragraphs that follow firstly examine a set of accuracy metrics[15] and then discuss the accuracy-implication evaluation through the two objectives of service level and stock-holding costs.

5.2.1 Accuracy Metrics

Accuracy evaluation exercises distinguish between the *in-sample* accuracy and the *out-of-sample* accuracy. The in-sample accuracy uses the historical data that are available and tries both to develop and to choose the forecast models' parameters using these data, typically by relying on the calculation of the estimated standard deviation of the forecast error in combination with the number of observations and the number of model parameters, e.g. the Akaike's Information Criterion (AIC). On the other hand, out-of-sample accuracy uses data

[15] The terms "metric" and "measure" are used interchangeably

that were not used for the development of the models. These data are used in order to evaluate the alternative forecast models.

The two approaches differ in their objectives. The in-sample approach aims to identify the best model/model parameters that are used for out-of-sample data forecasts. The out-of-sample approach then evaluates which model is the best in model evaluation/comparison studies. The analysis that follows assumes an out-of-sample requirement.

The metrics of forecast models' accuracy is an area of research that has received a lot of attention by the literature on forecasting, with the concerns being both theoretical and practical. Hyndman and Koehler (2006) correctly suggest that many of the known forecast accuracy metrics are not applicable to all cases. Furthermore, as Fildes (1992 p.85) stresses: "*Thus, the choice of error measures to summarise the error distribution should not merely be a question of personal preference, …, but rather, the forecaster must establish appropriate scaling and distributional assumptions for the data under analysis*".

The following list includes a summary of a number of attributes that the literature has pointed out as required for the accuracy metrics:

1. An accuracy measure needs to make sense and be easily understood by the decision makers (Armstrong and Collopy, 1992; Goodwin and Lawton, 1999; Kourentzes, 2013; Makridakis, 1993)
2. Equal sized positive and negative errors should be mapped on the same accuracy value (Goodwin and Lawton, 1999; Makridakis, 1993)
3. An accuracy measure needs to be "unit free" when it is used to compare methods that produce forecasts for diverse datasets, or as Makridakis (1993, p528) states it *"… otherwise, we compare apples and oranges in ways that make little sense*". Even though this requirement actually falls under number 1 above, it is better to consider it on its own because it has more often than not been used as being the same as the attribute of "scale-independence" that is discussed below (see e.g. Hoover (2006, p.34), or (Makridakis and Hibon, 2000)). The "unit free" attribute is not required to

be considered when the forecasting models are evaluated over the same single dataset

4. An accuracy measure needs to be robust in the following conditions:
    a. The accuracy measures need to be robust in their outputs when used with datasets which are different (Makridakis, 1993). This attribute is often called being “scale independent” (Hyndman, 2006). The problem to be dealt with is that those datasets that have comparatively large values might dominate the comparisons among the forecasting models (Armstrong and Collopy, 1992)

The approaches that have been used to deal with the previously described requirement of the accuracy measure to be “unit free”, have often been used at the same time to make the metric “scale independent”. Denominators are applied that aim both at the removal of the dimensional “units” and also to make the measure robust to the different size that the values can have between the datasets. These denominators are occasionally some polynomial function of the value A that is to be forecast, with the simplest of them being plain A like in the Absolute Percentage Error (APE) (Section 5.2.2.3). However, as shown later in the analysis (Section 5.2.2.3), they do not make the measure completely robust to scale differences.

There are also two other approaches that have been used in order to overcome the problem of having datasets with different scales. Firstly, the summarising of all the accuracy outputs can be done with a measure of location like the median which is insensitive to the actual values, and thus of the scales. Secondly, some measure functions use other denominators than the plain A that are not so dependent on the future value A:

- One such approach has been Makridakis' (1993) sAPE (symmetric Absolute Percentage Error). Approaches like this use a denominator that is a function of the future As. However, in this way these functions can still be influenced by the peculiarities of the future/out of sample values of the

datasets. The related issues are discussed more in the algebraic analysis that follows

- Other approaches use the error of another forecast model in the denominator. This forms a sort of comparison between the errors of the numerator's model and the denominator's making the ratio a relative error metric RAE (Relative Absolute Error). In these cases there are also variations in which some approaches use the out-of-sample data for their denominator's outputs, while others use the in-sample data

Nevertheless, the scale required independence attribute has been quite a challenge to the forecasting community both because it has not been fully met, but also because the suggested error functions are often not easy to be interpreted by the decisions makers (Kourentzes, 2013)

b. The possibility of division by zero should not exist (Makridakis, 1993). This is especially a challenge for intermittent datasets (Syntetos and Boylan, 2005)
c. The divisions by very large numbers (not just rare outliers) should not distort the output (Makridakis, 1993) by e.g. producing accuracy outputs that are hard to discriminate. In such cases, the differences among the errors are scored by the accuracy measure as minimal and thus, the measure is not able to distinguish among the competing models
d. Outliers in any existing dataset should not be influencing the evaluations of the forecasting models (Armstrong and Collopy, 1992; Makridakis, 1993)

What can be observed is that required attributes in 2, 3 and 4 are driven by the way that the forecast errors e have been embedded in the accuracy metric functions and also by the robustness/stability that the occasionally applied denominator can bring to the metric. Consequently, a study of the effects of e and especially of A in the denominator of the functions that use them could provide insight on the measures advantages and weaknesses.

Therefore, in order to study the different metrics and identify which are more applicable to the FPPs, it was necessary to investigate the functions formed to produce the metrics through algebraic analysis of their dependence on e and A. This was done by a series of algebraic methods which can be applicable to most of the forecast error functions.

The algebraic analysis was required since, as shown in the literature (Armstrong and Collopy, 1992; Davydenko and Fildes, 2016; Goodwin and Lawton, 1999; Koehler, 2001) the study of the functions' applicability and investigation of their potential weaknesses has mainly been done through empirical analysis, experience and intuition, while algebraic analysis is a more structured approach that can both prove literature's reported findings and potentially reveal others.

All of the functions studied below are evaluations of any *single* forecast. They take as input the error e of the forecast, treating it as the fundamental unit/argument, but also the value A that is to be forecast, and then by the use of the function $f(.)$, e and A map to a single value Y that expresses e's accuracy evaluation[16]: $e, A \xrightarrow{f(.)} Y = f(e, A)$. Consequently, the algebraic analysis of $f(.)$ studies how well this mapping serves the accuracy evaluations.

What should be kept in mind is that the algebraic study of such functions is on a *single* possible value and thus, it does not include statistical analysis. This means that the whole space domain of the values Y of the functions is examined irrespective from the fact that some of them may be less likely to be experienced than others. The results should then be considered along with the decision for the final metric.

The metric which is eventually used in order to evaluate a forecast model is a summary of a number of outputs from the functions that were studied and are presented below. The type of the summary to be used is usually chosen among one of the following measures of central tendency (Armstrong and Collopy, 1992):

[16] In the analysis further below, the notation of $Y = f(e)$ was changed into the more intuitive $AMF(e)$ for Accuracy Metric Function

- The arithmetic mean
- The geometric mean
- The median
- The trimmed mean
- The winsorized mean

5.2.2 Algebraic Analysis of the Accuracy Metric Functions

Most of the error measures are based on the L-step ahead forecast error:

$$e_T(L) = A_{T+L} - \hat{F}_T(L)$$

where T is the time when the forecast $\hat{F}_T(L)$ is produced for L-steps ahead, and A_{T+L} is the actual value when that future step comes at time $T + L$.

Using this simple metric, many different measures/functions have been developed by taking e.g. its absolute value, its squared value or the ratio of the square or of the absolute with other values. For brevity, any such function has been called Accuracy Metric Function, $AMF(e(A, \hat{F}), A)$ where again for brevity the $A_{T+L} = A$. Additionally, the error has sometimes been written as $e(A, \hat{F}) = e_T(L)$ to remind us that the error is itself a function of the actual future value A and its forecast $\hat{F}$. Furthermore, the error has been expressed as $e = e_T$ when there has been no need to investigate the relationship of the error to the actual future value and its forecast, while, given that in the FPP it is assumed that $\frac{Lead\ Time}{FPP\ period} > 1$ (Section 1.2) the L steps ahead have not been considered.

Furthermore, in discussing forecasting demand for spare parts, it has been assumed that the values of A are non-negative. In analytical terms, this assumption means that $A \epsilon \mathbb{N}$, i.e. that A is a countable number. However, since the objective is to develop an understanding of the general case, for the convenience of the calculations of limits and of derivatives it has been assumed that the generality is not reduced if A is considered as a non-negative real number, i.e. $A \epsilon \mathbb{R}_+$

The AMFs have been investigated using the following criteria:

1. Evaluate if the measure is definable for the whole domain of the function, i.e. for $\forall e \in \mathbb{R},\ A \epsilon \mathbb{R}_+,\ \hat{F} \in \mathbb{R}_+ \rightarrow\ AMF(e(A,\hat{F}),A) \in \mathbb{R}_+$. Observe the latter point that the AMF needs to be mapped to positive real values. The reason is that in order to use a measure of central location by taking the sum or the products of the AMF's outputs, these outputs need to have a positive sign.
2. Examine if the measure treats the positive and the negative error values equally, i.e. if for its whole domain, the $AMF\ \left(e\left(A,\hat{F}\right),A\right)$ is symmetric for the error (Goodwin and Lawton, 1999; Makridakis, 1993): $AMF(e(A,\hat{F}),A) = AMF(-e(A,\hat{F}),A)$
3. Study the $AMF(.)$ function according to $e\left(A,\hat{F}\right)$ and according to A
    a. This includes taking AMF's first and second derivative on e and on A, and comment on the outputs. Using these derivatives the aim is to facilitate the modeller's decision regarding her intentions about the evaluation. If the modeller wants to have outputs that are "fair" indicators, i.e. they produce outputs that treat higher error values the same as the lower, then the modeller should be looking for a constant first derivative, i.e. $\frac{d(AMF(e,A))}{de} = c$, with c a non-zero constant; in this case the second derivative does not provide any additional information. On the other hand, if the modeller wants something different, e.g. large errors to be penalised more, then the first derivative should be a positive, strictly monotonic function of the error, i.e. $\frac{d(AMF(e,A))}{de} = g(e) > 0$. The shape of this function $g(e)$ then needs to be further studied by the use of the second derivative
    b. Examine the behaviour of the AMF function through the whole domain of the errors e and the values A, which means that the limits of the function at the edges of e's and A's domains are estimated as well. Consequently, the function is calculated when the errors e / values A take extreme values, which for the present research cases they are $\pm\infty$, or make the AMF's denominator equal to 0

4. Discuss how the AMF and the resulting accuracy metric(s) can be interpreted by the decision maker

5.2.2.1 Squared Error Function (SE)

The Squared Error function is defined as:

$$AMF\big(e(A,\hat{F}),A\big) = AMF\big(e(A,\hat{F})\big) = e(A,\hat{F})^2$$

The actual value A in this case is only embedded inside $e(A,\hat{F})$.

1. Because $\forall e \in \mathbb{R},\ A\epsilon\mathbb{R}_+,\ \hat{F} \in \mathbb{R}_+ \rightarrow AMF\big(e(A,\hat{F})\big) = e(A,\hat{F})^2 \in \mathbb{R}_+$, there is no problem with the function to be used for any value of e, A or $\hat{F}$
2. $AMF\big(e(A,\hat{F})\big) = e(A,\hat{F})^2 = (-e(A,\hat{F}))^2 = AMF\big(-e(A,\hat{F})\big)$, so, by being symmetric both negative and positive errors are treated equally
3. The first derivative on the error is $\frac{d(AMF(e))}{de} = 2e, e \in \mathbb{R}$, which means that the error values are not mapped to equally changing values. Furthermore, $\frac{d^2(AMF(e))}{de^2} = 2 > 0$. This means that as the error values get further away from 0, the resulting accuracy metric function's values change faster, and thus the higher errors are penalised more. (These results are actually the results that one would typically get from a parabola concave upwards centred on 0).

   Now, in order to see the effect that the values of A can have on the shape of the AMF functions and its derivatives, the function is written as follows:

   $AMF\big(e(A,\hat{F})\big) = (A-\hat{F})^2$, $\frac{d(AMF(e))}{de} = 2\big(A-\hat{F}\big)$ and $\frac{d^2(AMF(e))}{de^2} = 2$. Both the function and its first derivative on the error depend on the actual value A.

   The Squared Error function treats different magnitudes of As in a different way, and this can be seen through the function itself and its first derivative. Firstly, this $AMF(.)$ function penalises the errors differently, as shown by the dependence of the plain function on A. Secondly, the rate of change in the penalisation changes as the error values get higher/lower with different datasets/As. This is shown by the fact that the first derivative is linearly

dependent on A. Theoretically though, neither of the two problems happen if the models' forecasts $\widehat{F}$ can keep the difference from A the same in all datasets; however, this is not always likely.

Furthermore, if for example there are two datasets in which one has a higher spread in its data than the other, then the combined evaluation is more affected by that more dispersed dataset. This is something that has been reported in the literature as well. For example Thompson (1990) and Armstrong and Collopy (1992) suggest that unless there are many datasets to compare, taking the Mean of the Squared Error (MSE), or equivalently taking the Root of the Mean ($RMSE$) gives an unreliable indicator of the accuracy.

Similarly, in cases the measure is used for multimodal datasets, then the accuracy evaluation through the Squared Error can become challenging.

Finally, taking the limits of the errors e on $\pm\infty$, or on 0 does not add to the study.

4. Given the effect that A can have on the SE, the $RMSE$ is suitable only in the case where all forecast models are evaluated against a single dataset, while in such situations the interpretation to decision makers is not very straightforward (Armstrong and Fildes, 1995)

5.2.2.2 Absolute Error Function (AE)

The Absolute Error function is defined as:

$$AMF\left(e(A,\hat{F}),A\right) = AMF\left(e(A,\hat{F})\right) = \left|e(A,\hat{F})\right|$$

Again, the actual value A is only embedded inside $e(A,\hat{F})$.

1. For $\forall e \in \mathbb{R},\ A\epsilon\mathbb{R}_+, \widehat{F} \in \mathbb{R}_+ \rightarrow\ AMF\left(e(A,\hat{F})\right) = \left|e(A,\hat{F})\right| \in \mathbb{R}_+$
2. $AMF\left(e(A,\hat{F})\right) = \left|e(A,\hat{F})\right| = \left|-e(A,\hat{F})\right| = AMF\left(-e(A,\hat{F})\right)$, the function is symmetric in regards to e, so negative and positive errors are treated equally

3. For $\frac{d(AMF(e))}{de} = 1,\ e \in \mathbb{R}^{+}$ and $\frac{d(AMF(e))}{de} = -1,\ e \in \mathbb{R}^{-}$and also that $\frac{d^2(AMF(e))}{de^2} = 0$. This means that the error values are mapped into equally changing AMF values which are first reducing for the negative errors until they reach zero and then increasing for the positive errors. Expressed in a different way, as the error values get further away from 0, the function linearly departs from the origin $O(0,0)$ having a fixed constant 45^oangle with the Oy axis. These results are actually the results that one would typically get from an absolute linear function with a slope of 1 and centred on 0.

   Now, in order to see the effect that the values of A and $\widehat{F}$ can have on the shape of the $AMF(.)$, it is written as follows:

   $AMF\left(e(A,\widehat{F})\right) = \left|A - \widehat{F}\right|$. The function depends linearly on the actual value A and its forecast $\widehat{F}$. Again, just like the Squared Error function, with the Absolute Error function if there are different datasets, then they could also have different magnitudes of A. Consequently, since this $AMF(.)$ depends on A, it could penalise the errors differently among datasets. On the other hand, for a single dataset the rate of change in the penalisation as the error values get higher is unchanged as shown by the fact that the first derivative is independent of A.

   Finally, taking the limits of the errors e on $\pm\infty$, or on 0 does not add to the study.

   Using the mean of a number of the AE outputs gives the widely used Mean Absolute Error (MAE) metric (also called Mean Absolute Deviation (MAD)) which has the dimensions of the forecast. As compared to MSE this does not have the problems of treating differently the different changes in the values of the error. However, even though to a lesser extent than MSE, due to its function MAE still gets affected by the datasets with very different A values (Syntetos and Boylan, 2005).

   An additional challenge of the MAE that has been reported (see e.g. Davydenko and Fildes (2016)) is that since the Absolute Error function

turns all errors to become non-negative, its outputs tend to be skewed to the right and then the applied mean is not representative.

4. MAE's interpretation is straightforward. It is the average absolute deviance of the out-of-sample data from each forecast model's predictions

5.2.2.3 Absolute Percentage Error Function (APE)

The Absolute Percentage Error function is a widely used accuracy measure function, often due to its *intended* objective to be scale-independent:

$$AMF(e(A), A) = AMF(e(A)) = \left|\frac{e(A)}{A}\right|$$

However, its robustness has often been debated (Goodwin and Lawton, 1999; Makridakis, 1993), and as is shown by the algebraic analysis, it does depend on the scale when the values of A are extreme (very large or very low) relative to the errors e.

1. For $\forall e \in \mathbb{R},\ A \epsilon \mathbb{R}_+^*,\ \hat{F} \in \mathbb{R}_+ \rightarrow\ AMF\left(e(A, \hat{F})\right) = \left|\frac{e(A,\hat{F})}{A}\right| \in \mathbb{R}_+$.

   However, if:

   a. $limA = 0^+$ then it is either

      $AMF\left(e\left(A, \hat{F}\right), A\right) \rightarrow +\infty, e(A) \in \mathbb{R}^*$, or

      $AMF\left(e\left(A, \hat{F}\right), A\right)$, the function is undefined when $e(A) = 0$ (Hyndman and Koehler, 2006)

      The above state two things. Firstly, and as many researchers have commented (Boylan and Syntetos, 2006; Hoover, 2006; Hyndman, 2006; Makridakis, 1993), the measure cannot be defined if in the dataset there can be $A = 0$, which is a significant problem when the data demonstrate an intermittent behaviour. Secondly, the measure produces really high, undiscriminating values if, as compared to A, the forecast errors are very large. This is demonstrated in Figure 5-2. As As get smaller, the values of APE jump from one curve to the other at steps of very different size even for very neighbouring As. Furthermore, the problem gets larger at the lower areas of As

relative to the values of the error. More details are given in step 3 below

b. $limA = +\infty$, then:

$AMF\left(e\left(A,\hat{F}\right),A\right) \to 0, e(A) \in \mathbb{R}_{\pm}^{*}$

The observation here is similar to the one stated above, and which seems to have been missing from the literature. If the forecast errors are very small as compared to A, then the accuracy measuring function is producing really small, undiscriminating values. Again, this is demonstrated in Figure 5-2. As As get larger the values of APE are not very different to each other, regardless the size of the errors (the different curves). More details are given in step 3 below

Furthermore, the ratio of the e over A has to be definable, i.e. the values to be referenced to a clearly defined zero (Hyndman, 2015), that is to be of a "ratio" and not of an "interval" type. This is not a problem in the uses that have been applied in the present research, because the models forecast the mean number of demands for spares where there is a meaningful/"absolute" zero set to "no demands on average".

2. $AMF\left(e\left(A,\hat{F}\right),A\right) = \left|\frac{e(A,\hat{F})}{A}\right| = \left|-\frac{e(A,\hat{F})}{A}\right| = AMF\left(-e\left(A,\hat{F}\right),A\right)$, so, in the areas where the function can be defined , negative and positive errors are treated equally
3. For $\frac{d(AMF(e,\hat{F}),A)}{de} = \frac{1}{A}, e \in \mathbb{R}^{+}$ and $\frac{d(AMF(e,\hat{F}),A)}{de} = -\frac{1}{A}, e \in \mathbb{R}^{-}$, with $A \in \mathbb{R}^{*}$ and also $\frac{d^2(AMF(e,\hat{F}),A)}{de^2} = 0$. Like in the AE earlier, these outputs indicate that the error values are mapped into equally changing values, but which, in the case of APE, do depend on A. This means that as the error values get further away from 0, the function's values linearly depart from the origin $O(0,0)$ having an $\arctan\left(\frac{1}{A}\right)$ angle with the Oy axis, which is not fixed since it depends on A. These derivatives also suggest that if the set of forecast errors are concentrated around zero, the mass of the APE's outputs are

also close to zero, having a long tail of the few positive values away from zero. The result will then be to have the distribution of the APEs right-skewed and asymmetrical (Boylan and Syntetos, 2006). However, the distribution's a/symmetry attribute refers to the mass of the values. Asymmetry in this case should not be confused with the function's a/symmetry that is studied here and which refers to the a/symmetry of the values that the Accuracy Metric Function can take per symmetrical pair of vales of its arguments.

Actually, the linear departure of the APE's values from the origin *could* have been a desirable attribute. Apart from making the measure "unit free", the reason for introducing the value A in the denominator is to have a rate of change in the penalties to errors e that are proportional to A and in this way to also make the function "scale-independent". However, as stated earlier as well, comparatively really high/low values of A tend to make the rate of change of the Accuracy Metric Function unusably extreme. In support of this observation, one can find in the literature debates like the one discussed right after, about As' intended and actual effects on the AMF.

The discrepancy as presented above, has been brought as an example by Makridakis (1993), but has not been highlighted or generalised as it is done next here. In his analysis, Makridakis compared the APE output of a forecast of 100 when the actual value A is 150 (absolute error of 50) which gives $APE = \frac{|150-100|}{150} = 0.33$, to a forecast of 150 when the actual value A is 100 (absolute error of 50 again) which gives a different value $APE = \frac{|100-150|}{100} = 0.50$. In their comment Goodwin and Lawton (1999) object to this statement of asymmetry but do not investigate the root cause that Markidakis highlighted:

$$APE_1 = \frac{|e|}{A_1} \neq \frac{|e|}{A_2}$$

, and the important observation in the above is that the difference in the output for the same amount of $|e|$ takes place for *any* $A_1 \neq A_2$, however little distance apart the two values might have. In essence, what Markidakis stated is that for the same absolute error, the changes in the APEs in close but different As are unequal, and even more, as the first derivative shows this inequality changes magnitude as the values of A change. To see that in more details, it is first required to analyse the AMF from the A's perspective.

Taking the derivatives of the function with reference on A gives:

$$\frac{d(AMF(e,\hat{F}),A)}{dA} = -\frac{|e|}{A^2} < 0, \forall A \in \mathbb{R}_+^* \quad \textbf{[5-1]}$$

$$\frac{d^2(AMF(e,\hat{F}),A)}{dA^2} = 2\frac{|e|}{A^3} > 0, \forall A \in \mathbb{R}_+^* \quad \textbf{[5-2]}$$

These results imply that for the same amount of (absolute) error $|e|$, the penalisation reduces with the square of the datapoint's value A, while the second derivative shows that the curvature of the function remains concave up (reducing fast) for the whole domain of $A \in \mathbb{R}_+^*$. This observation suggests a great sensitivity to the values of A – especially the low – making the function problematic for an intended use to be a scale-independent error measure, especially among datasets in which some of them include really low values.

The observation above suggests the existence of the following two problems. In the lower values the changes in the penalties can be very dramatic between neighbouring values of A, while in a series of high values if one non-high value is included it can determine the result if additive location measures like the mean are used.

Figure 5-2 provides a complementary view of this problem. Starting for example with a value of $A = 2$ and then going vertically up, the cuts/output values of the different curves at $APE(e, A = 2)$ are of certain equal differences among them: for the error values in the figure $\Delta(APE(e, A =$

$2) = \frac{1}{2}$. However taking half a unit to the left twice i.e. for $A = 1.5$ and $A = 1$, and again going vertically up, in both the two new cuts the differences are equal within each curve, but they are not the same among the curves: again for the error values in the figure $\Delta(APE(e, A = 1.5) = \frac{2}{3}$, and $\Delta(APE(e, A = 2) = 1$. These differences are not linearly proportional to each other. In other words, the changes in the penalty are not linearly proportional to the changes in the values of A, and this problem increases as the values of A get lower.

Furthermore, taking any of the curves, the higher the values of A become then the less the APE values get different from each other, and thus the function gets less effective in differentiating among the errors.

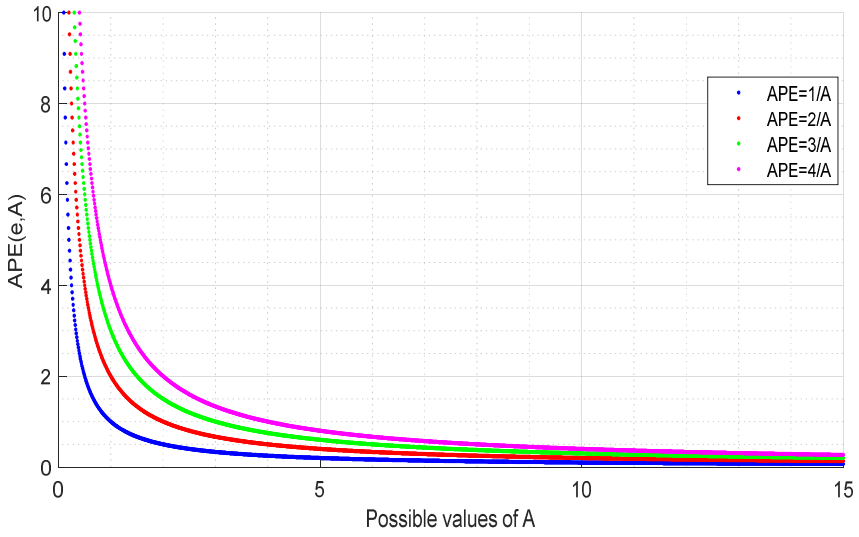


Figure 5-2: APE values for a range of possible values of A. Each curve is for a different absolute error $|e| = 1, 2, 3, 4$

The reason for both of these observations is APE's rate of change according to A which is proportional to $\frac{1}{A^2}$ and which of course is also related to the limits that were examined in 1a and 1b above.

One final point that highlights the importance of this observation and also of the algebraic analysis method, has to do with the suggestion of still using the APE when $A = 0$, by replacing it with $A = 1$ and thus get the

Denominator-adjusted MAPE (DAM) (Hoover, 2006). Using intuition only, one could say that it could be a practical solution. However, algebraic analysis in such a case raises questions that can challenge such a decision. Using DAM, the function's outputs of infinity $\frac{|e|}{A} \to +\infty$, are replaced by $|e|$, even if the value of $|e|$ is small compared to other errors that correspond to non-zero As.

The root cause of the above lies in the intended use of A as the AMF's denominator. The denominator is introduced in order to make the Accuracy Metric Function of $|e|$ both unit and scale independent. While unit independence is indeed (easily) accomplished, scale independence is not. The measured error is about the forecast of the location parameter of the distribution that models the As, and not of the value A iself. Individual As can fall in the areas that could be outliers, or belong to the tails of highly skewed datasets and thus be far from the (unknown) true location parameter. Using these As instead of the referenced scale does not eventually work as a reference/ratio and makes the resulting AMF sensitive to their peculiarities.

As Fildes (1992, p.83) states the function should not depend on the datasets and "*any automatic forecasting system should not be geared to respond to such extremes*", while he also calls their effects on the function "*contamination*". The APE function is always "noisy" because it is sensitive due to A, or as Fildes (1992, p.85) observes "… *MAPE, MdAPE etc. … depend to a greater or lesser extent on* [outliers]. *Equally important, all measures based on* APE *suffer from the lack of equivalence across series and across time*[17]". Consequently, the modeller needs to decide if he/she is content with the amount of discrepancies or not, but he/she should be cautious if in the dataset(s), as compared to the forecast errors, there are either many low values of A, or many high values.

[17] This is the same as saying that the measures based on APE can provide different outputs from record to record in a single dataset and among different datasets

Using the mean of a number of APE outputs gives the widely used Mean Absolute Percentage Error (MAPE) metric, while, if the concern is if *occasional* outliers might affect the forecast models' evaluations, then the use the median provides the also widely used MdAPE. However, even in this case the distributional asymmetry along with the need to preserve the information provided by outliers, has led a number of researchers to suggest that when APE produces outliers, to firstly transform them to make the APE more stable (Coleman and Swanson, 2007; Davydenko and Fildes, 2016; Swanson, Tayman, and Barr, 2000).

4. Despite the occasional debates, MAPE is still being widely suggested when the datasets are not prone to the problems discussed above (Boylan and Syntetos, 2006). MAPE is also quite simple to calculate and intuitive when presented to a decision maker since it is understood as the average percentage error that the forecast model produces over the values that can be experienced

5.2.2.4 Symmetric Absolute Percentage Error Function (sAPE) – variant 1

The symmetric Absolute Percentage Error loss function (Makridakis and Hibon, 2000) was introduced in order to deal with the problem of APE's asymmetry as highlighted in (Makridakis, 1993), and variant 1 has been discussed in a number of papers (see e.g. (Hyndman, 2006)):

$$AMF\left(e\left(A,\widehat{F}\right),A,\widehat{F}\right)=\frac{|e(A,\widehat{F})|}{\frac{A+\widehat{F}}{2}}$$

This measure seems to be able to handle cases in which A is zero or close to that value (Hryniewicz and Kaczmarek, 2016; Hyndman and Koehler, 2006; Sujjaviriyasup, 2017), but as it is shown in 1b below, it is not done very effectively.

Next, the algebraic analysis is used in order to identify the problems that might be faced when using variant 1 of $sAPE$ as a loss function.

1. For $\forall e \in \mathbb{R},\ A\epsilon\mathbb{R}_+,\ F \in \mathbb{R}_+$ the $AMF\left(e\left(A,\widehat{F}\right),A,\widehat{F}\right)=\frac{|e(A,\widehat{F})|}{\frac{A+\widehat{F}}{2}}$. In order to be able to study the function's domain according to e and A the AMF is transformed as follows:

$AMF(e,A) = \frac{|2e|}{A+\hat{F}+A-A} = \frac{|2e|}{2A-e}$, which is definable in $\mathbb{R}$ (*and not just* $\mathbb{R}_+$) for $\forall e \in \mathbb{R}\backslash 2A,\ A \epsilon \mathbb{R}_+ \backslash \frac{e}{2}$, the latter considered only when $e > 0$ under the applied assumption that A is non-negative.

Firstly, the above demonstrates Hyndman and Koehler's (Hyndman and Koehler, 2006) point that even though it is called "absolute", in this form it is not definable only on $\mathbb{R}_+$ as it is required for an accuracy measure loss function (see earlier discussion in the Introduction), but it can also take negative values when $e > 2A$. However, examining the error's requirement for negativity $e > 2A \Rightarrow A - \hat{F} > 2A \Rightarrow 0 \geq -A > \hat{F}$ shows that in this specific variant of sAPE the negative values would be experienced if the forecasts $\hat{F}$ are not just negative, but also lower than $-A$, something should not be expected from models that try to forecast for distributions of non-negative values $A \geq 0$.

In order to study the structure of the function, it is written as follows:

$$AMF(e,A) = \begin{cases} \frac{2e}{2A-e}, e \geq 0, e \neq 2A, A \geq 0 \\ \frac{-2e}{2A-e}, e < 0, A \geq 0 \end{cases} \quad \textbf{[5-3]}$$

a. Looking at the function's limits close to the point where it is not defined for e, the following results are acquired[18]:

$\lim_{e \to 2A^+} AMF(e,A) = -\infty$, $\lim_{e \to 2A^-} AMF(e,A) = +\infty$

, which suggests that in the area of $2A$, the situation is similar to what APE had close to 0, but in this case the problems are not as easily identifiable because they depend on A, and even more, for the present $sAPE$ variant they jump from $+\infty$ to $-\infty$. On the other hand, as it was discussed earlier, this phenomenon takes place for non-positive forecasts, which is something that should not be expected for non-negative datasets.

[18] Observe that since the discussion is about the limits around $2A > 0$, both these limits correspond to the upper branch of the AMF in**Error! Reference source not found.**

b. Moreover, investigating for the different values of A:

$\lim_{A\to+\infty} AMF(e,A) = 0$, while for $A = 0$ the $AMF(e, A = 0) = 2$ for any negative error $e \in \mathbb{R}^{*}_{-}$ and $AMF(e, A = 0) = -2$ for any positive error $e \in \mathbb{R}^{*}_{+}$ which is the result also reported in (Boylan and Syntetos, 2006; Makridakis and Hibon, 2000).

c. If the forecast $\widehat{F}$ is exactly the same as the value A, then the error e and the $sAPE$ are equal to 0

d. If the forecast $\widehat{F}$ is equal to 0 and thus the error e is equal to A, then from the upper branch of **[5-3]** the function's value becomes equal to 2. This penalty is twice as much as the respective APE's

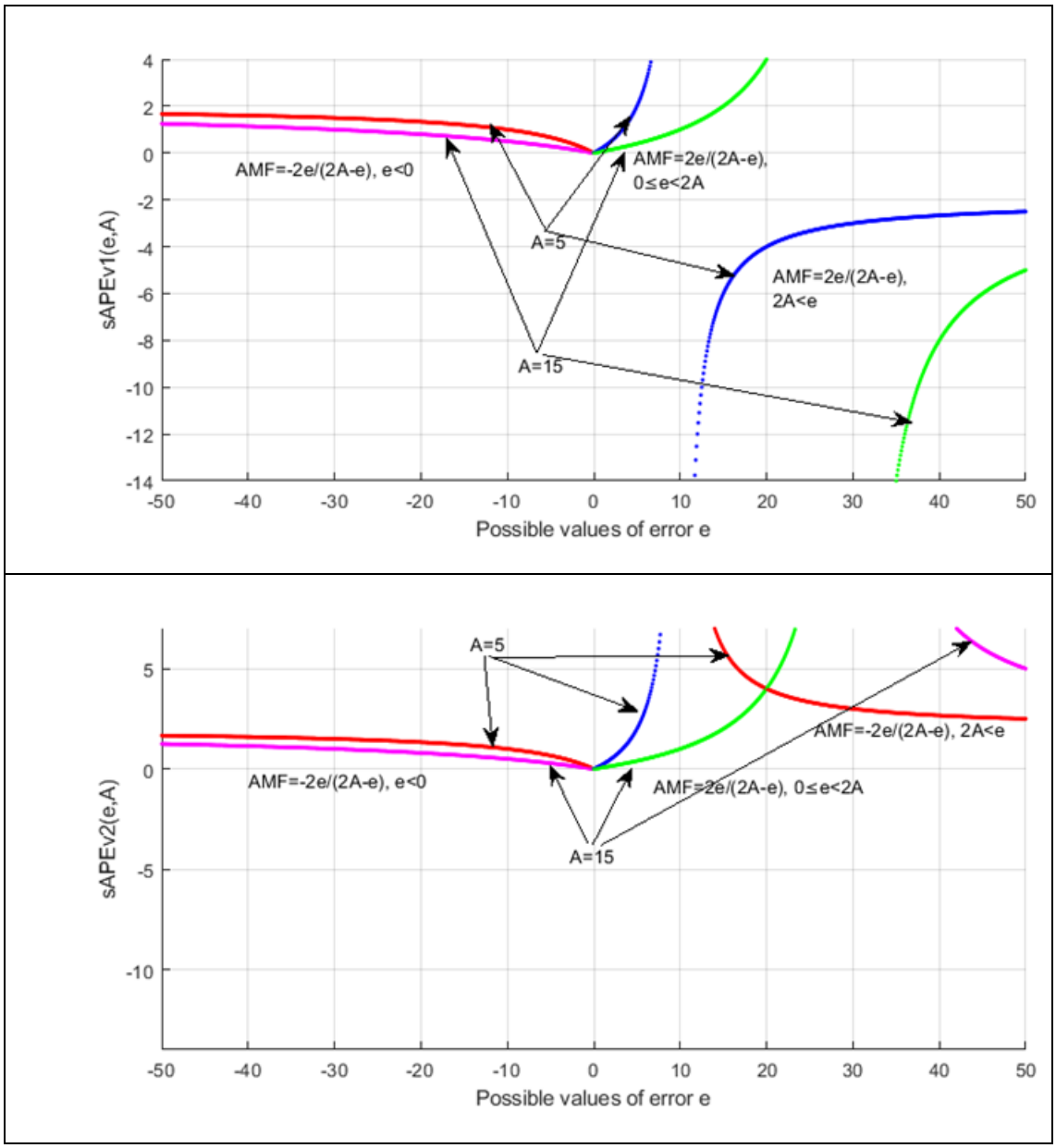
e. Additionally, taking the error limits to infinity, gives:

$\lim_{e\to-\infty} AMF(e,A) = 2$, $\lim_{e\to+\infty} AMF(e,A) = -2$

2. Without loss of generality, in order to investigate the symmetry in relation to zero, a value r for the error is chosen so that $e = r \geq 0$. Then:

$AMF(e = r, A) = \frac{2r}{2A-r} \neq AMF(e = -r, A) = \frac{-2(-r)}{2A-(-r)} = \frac{2r}{2A+r}$, and thus, the examined variant of sAPE is not symmetric in relation to zero.

Additionally, one point that this analysis also shows and which has not been included in the relevant studies that the author has found in the literature, is that sAPE is also not symmetric because it is defined for $e = -2A < 0$ but not for $e = 2A > 0$ as it can also be observed in Figure 5-3 below. However, as it was again mentioned earlier, the latter stands for an error that would come from a non-positive forecast $\widehat{F} = -A < 0$


sAPEv1(e,A)
Possible values of error e
AMF=-2e/(2A-e), e<0
AMF=2e/(2A-e),
0≤e<2A
A=5
AMF=2e/(2A-e),
2A<e
A=15
sAPEv2(e,A)
Possible values of error e
A=5
AMF=-2e/(2A-e), 2A<e
AMF=-2e/(2A-e), e<0
AMF=2e/(2A-e), 0≤e<2A
A=15

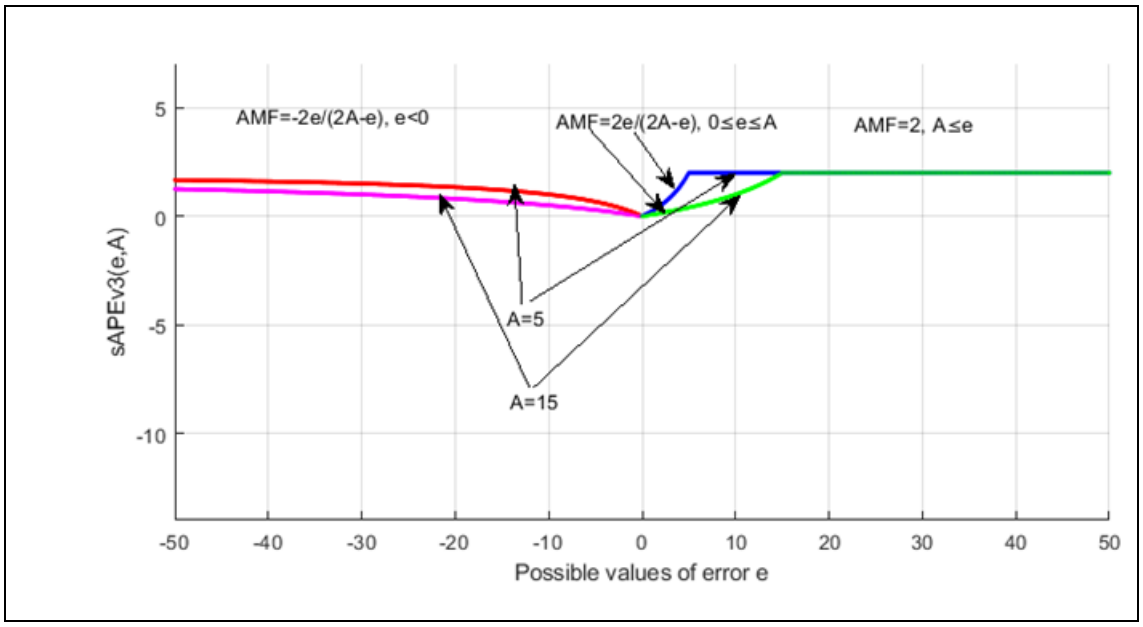


Figure 5-3: $sAPE$ v1, v2 and v3 for a number of possible values of e. Each line is for $A = 5, 15$. Observe that all three variants are exactly the same for $-\infty \leq e \leq A$

3. The function's first and second derivatives on the error e are as follows:

$$\frac{d(AMF(e,A))}{de} = \begin{cases} \frac{4A}{(2A-e)^2}, e \geq 0, e \neq 2A, A \geq 0 \\ \frac{-4A}{(2A-e)^2}, e < 0, A \geq 0 \end{cases}$$

$$\frac{d^2(AMF(e,A))}{de^2} = \begin{cases} \frac{8A(2A-e)}{(2A-e)^4}, e \geq 0, e \neq 2A, A \geq 0 \\ \frac{-8A(2A-e)}{(2A-e)^4}, e < 0, A \geq 0 \end{cases}$$

There are a number observations that can be made from the above. Firstly, if $A = 0$ the first derivative will be zero for any error value e within its domain. This sheds more light to the earlier observation that the AMF's value will be ± 2 without any tendency to change apart from when the error is also zero where the AMF is undefinable. This situation is also presented in Figure 5-4.

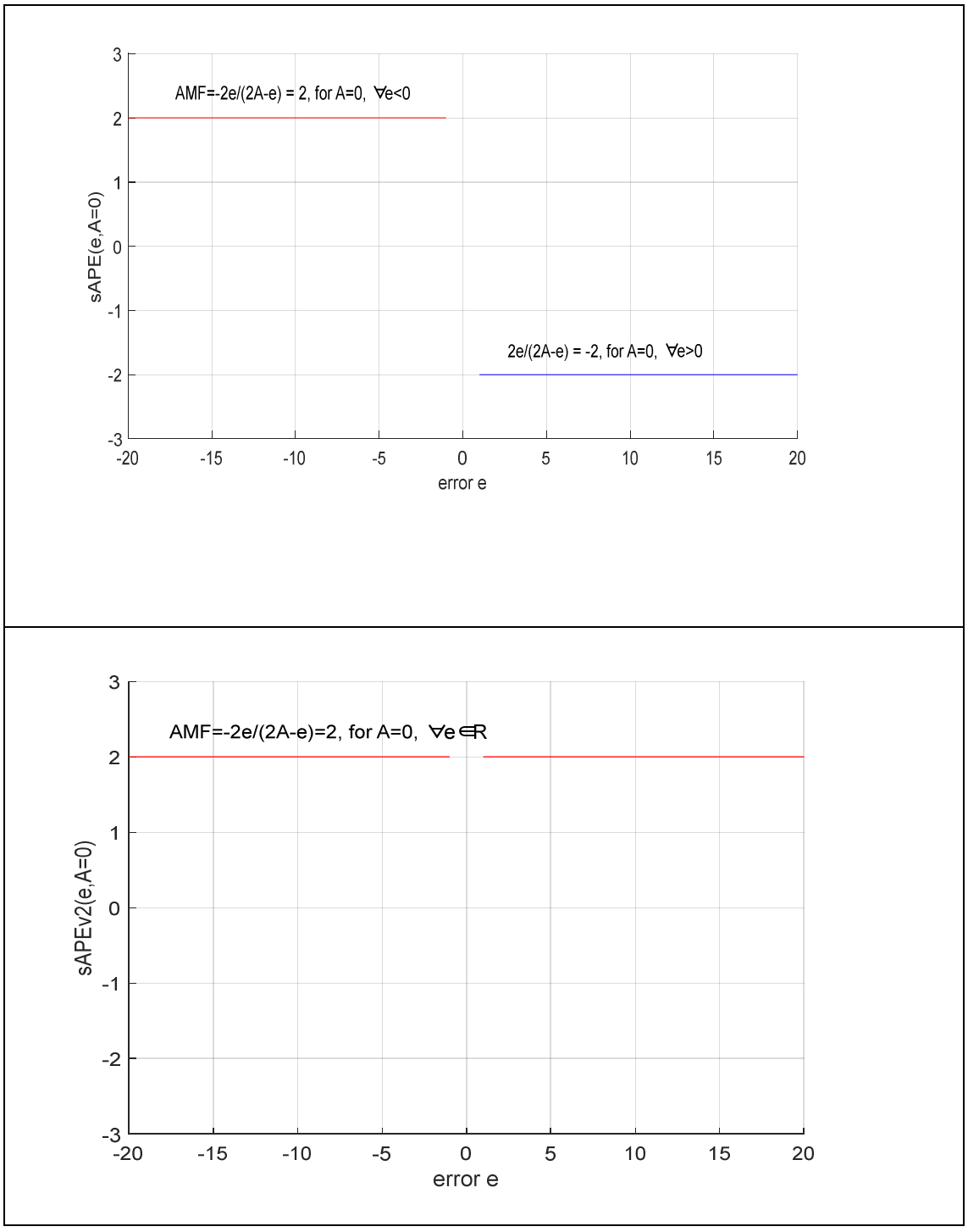


Figure 5-4: Plots of $sAPE$v1 and $sAPE$v2 when $A = 0$. $sAPE$v3 is identical to $sAPE$v2

Secondly, starting from $e < 0$, for the very large negative values of e (which are also large in their absolute values) the AMF has values close to $+2$ (see also Figure 5-3, (Goodwin and Lawton, 1999, fig.1)) and similar practical discussions

in (Hyndman and Koehler, 2006; Koehler, 2001)). Furthermore, its first and second derivatives are always < 0 which means that the AMF values decrease while the curve's shape is concave down. Moreover, as it can be seen by the values of the first derivative, this decrease happens slowly - which implies a small discriminating ability - and from $+2$ it reaches 0 as e gets closer to zero where the derivative gets its maximum value. The zero error value is also the local minimum of $AMF = 0$.

From then on, $e > 0$ and the upper branch of [**5-3]** takes over. Both the first and the second derivatives become positive, the curvature changes and the AMF increases rapidly and from 0 it tends to $+\infty$ as $e \to 2A^-$ (Koehler, 2001). For $e = 2A$ the AMF is not defined, but as $e \to 2A^+$the AMF starts from $-\infty$. The first derivative is still positive which means that the function still increases its values (from the $-\infty$ that it starts) while the second derivative becomes negative and thus the curvature changes once more which means that the rate of increase reduces rapidly, but the AMF never becomes larger than -2 which is its limit as $e \to +\infty$.

Using the assumption that the forecasts $\hat{\mathrm{F}}$ are non-negative, one can also see which parts of the AMF correspond to which values of such forecasts range (see also Figure 5-3). The left part of the figure up to $e = 0$, corresponds to the lower branch of [**5-3]** and it is for $e < 0 \Rightarrow 0 < A < \widehat{F}$. The right, positive part of the figure corresponds to the upper branch. Using the requirement that the denominator and the error should both be positive so that $sAPE$ is on its positive range, it should be that $2A - e > 0 \Rightarrow A > -\widehat{F} \Rightarrow \widehat{F} > -A \Rightarrow \widehat{F} > 0$, and $e \geq 0 \Rightarrow A \geq \widehat{F}$. These two results imply that any over-forecasting is mapped by the left part of the figure and the under-forecasting by the right part up to $\widehat{F} = A$, that is the AMF does not reach the really large values close to $2A^-$.

5.2.2.5 Symmetric Absolute Percentage Error Function (sAPE) – variant 2

This variant of the symmetric Absolute Percentage Error loss function is the one that was initially introduced by Makridakis (Makridakis, 1993):

$AMF\left(e\left(A,\widehat{F}\right),A,\widehat{F}\right)=\left|\frac{e\left(A,\widehat{F}\right)}{\frac{A+\widehat{F}}{2}}\right|$

and is still used by researchers (Hamza et al., 2018)

The algebraic analysis that follows aims to investigate the problems that might be faced when using this variant of the $sAPE$.

1. For $\forall e\in\mathbb{R},\ A\epsilon\mathbb{R}_+,\ \widehat{F}\in\mathbb{R}_+$ the $AMF\left(e\left(A,\widehat{F}\right),A,\widehat{F}\right)=\left|\frac{e\left(A,\widehat{F}\right)}{\frac{A+\widehat{F}}{2}}\right|$. In order to be able to study the function according to e and A the function is transformed as follows:

$AMF(e,A)=\left|\frac{2e}{A+\widehat{F}+A-A}\right|=\left|\frac{2e}{2A-e}\right|$, which is definable in $\mathbb{R}_+$ for $\forall e\in\mathbb{R}\backslash 2A$, $A\epsilon\mathbb{R}_+\backslash\frac{e}{2}$, the latter considered only when $e>0$.

In order to be able to investigate what happens to the function at the limits of its domain, and also to see if it is symmetric and be able to take the derivatives, the function is written as follows:

$$AMF(e,A)=\begin{cases}\frac{2e}{2A-e},0\le e<2A,(e\neq 2A),A\ge 0\\ \frac{-2e}{2A-e},e<0\vee 2A<e,(e\neq 2A),A\ge 0\end{cases} \quad \textbf{[5-4]}$$

- For the function's limits close to the point where it is not defined, the present AMF gives the following outputs[19]:

$\lim_{e\to 2A^+}AMF(e,A)=+\infty$, $\lim_{e\to 2A^-}AMF(e,A)=+\infty$ a result which is different than the previously examined variant of sAPE, and which is also expected because this variant can take only non-negative values. Nevertheless, these outputs suggest again that around the area of $2A$, the problems are similar to what APE had close to zero, and again these problems are not as easily identifiable because they

[19] Observe that since the discussion is about the limits around $2A$, the left side corresponds to the upper branch of [**5-4**], while the right side corresponds to the lower branch

depend on A. However, as it was discussed earlier for the variant 1, this can only take place for non-positive forecasts.

- For the edges of A's domain:

$\lim_{A\to+\infty} AMF(e,A) = 0$, for all error values e, while for $A = 0$ the function gives again what Boylan and Syntetos report, i.e. $AMF(e, A = 0) = 2$ for any non-zero error ($e \in \mathbb{R}^*$) (Boylan and Syntetos, 2006)

- If the forecast $\widehat{F}$ is exactly the same as the value A, then the error e and the $sAPE$ are equal to 0
- If the forecast $\widehat{F}$ is equal to 0 and thus the error e is equal to A, then from the upper branch of [**5-4**] the function's value is equal to 2. This penalty is twice as much as the respective APE's
- Additionally, if the error limits are taken to infinity, the function gives:

$\lim_{e\to-\infty} AMF(e,A) = 2$, $\lim_{e\to+\infty} AMF(e,A) = 2$

2. Without loss of generality, in order to investigate the symmetry around zero, a value r for the error is again chosen so that $e = r > 0$ with:
   - $0 \le r < 2A, (r \neq 2A), A \ge 0$. This means that $-r < 0$, so while r is mapped using the upper branch, $-r$ is using the lower branch of [**5-4**] :

$AMF(e = r, A) = \frac{2r}{2A-r} \neq AMF(e = -r, A) = \frac{-2(-r)}{2A-(-r)} = \frac{2r}{2A+r}$, which shows that for $-2A < e < 2A$ this variant of sAPE is not symmetric in reference to zero.

   - $2A < r, (r \neq 2A), A \ge 0$. Then $-r < -2A < 0$. For both r and $-r$ correspond to the lower branch of [**5-4**]:

$AMF(e = r, A) = \frac{-2r}{2A-r} \neq AMF(e = -r, A) = \frac{-2(-r)}{2A-(-r)} = \frac{2r}{2A+r}$, which shows that for the range $e < -2A \wedge e > 2A$ this variant of sAPE is not symmetric there either

   - Furthermore, as before, sAPE is also not symmetric because sAPE is defined for $e = -2A < 0$ but not for $e = 2A > 0$ as it can also be observed in Figure 5-3

3. The function's first and second derivatives on the error e are as follows:

$$\frac{d(AMF(e,A))}{de} = \begin{cases} \frac{4A}{(2A-e)^2}, 0 \le e < 2A, (e \neq 2A), A \ge 0 \\ \frac{-4A}{(2A-e)^2}, e \le 0 \vee 2A < e, (e \neq 2A), A \ge 0 \end{cases}$$

$$\frac{d^2(AMF(e,A))}{de^2} = \begin{cases} \frac{8A(2A-e)}{(2A-e)^4}, 0 \le e < 2A, (e \neq 2A), A \ge 0 \\ \frac{-8A(2A-e)}{(2A-e)^4}, e \le 0 \vee 2A < e, (e \neq 2A), A \ge 0 \end{cases}$$

Again, if $A = 0$ the first derivative will be zero for any error value e. The situation is also presented in Figure 5-4 and the rate of change related problems are similar to what was discussed for variant 1. The largest difference is for $e > 2A$. In that area, the loss function is positive but the penalties are reducing as the errors increase.

Finally, if [**5-3]** is compared to [**5-4**], or the two top graphs of Figure 5-3 are compared, for the errors' range from -∞ to A, the two variants of sAPE are identical.

5.2.2.6 Symmetric Absolute Percentage Error Function (sAPE) – variant 3

According to (Hyndman, 2014), this variant of the symmetric Absolute Percentage Error loss function has firstly appeared in a working paper (Chen and Yang, 2004) and is used in a number of forecast modelling packages (Spider Financial, 2018):

$$AMF\left(e\left(A,\widehat{F}\right),A,\widehat{F}\right) = \frac{2|e(A,\widehat{F})|}{|A|+|\widehat{F}|}$$

Furthermore, most of the recent papers that the present work examined in order to see the $sAPE$ variants' applications, use variant 3 ((Andrawis and Atiya, 2009; Boulkaibet et al., 2017; Cavdar and Aydin, 2015; Liu et al., 2018; Marcot et al., 2006; Martínez-Álvarez et al., 2015; Štěpnička et al., 2013; Valle Dos Santos and Vellasco, 2015; Zamora-Martínez et al., 2013)). Nevertheless, just like in variant 2, the absolute values in the denominator give it the advantage of not producing negative values as the variant 1 does. Like variants 1 and 2, this variant seems to be able to handle cases in which A is zero, but again ineffectively.

For variant 3, $\forall e \in \mathbb{R},\ A \epsilon \mathbb{R}_+,\ \widehat{F} \in \mathbb{R}_+$ and $AMF\left(e\left(A, \widehat{F}\right), A, \widehat{F}\right) = \frac{2|e(A,\widehat{F})|}{|A|+|\widehat{F}|}$. In order to be able to study the function according to e and A the function can be transformed as follows: $AMF(e, A) = \frac{2|A-\widehat{F}|}{|A|+|\widehat{F}|} = \frac{2|A-\widehat{F}|}{A+|\widehat{F}|}$.

For $\widehat{F} \in \mathbb{R}_+$, the AMF becomes $AMF(e, A) = \frac{2|e|}{A+F}$. This form of the function when transformed to include only the error e and the value A, becomes exactly the same function as [**5-3]** and [**5-4**] of variants 1 and 2, which means that the three variants produce exactly the same outputs for errors less than A. This observation can also be seen by comparing the plots in Figure 5-3. Furthermore, this variant 3 reduces the previous variants' problem of getting really high values in the area of e around $2A$, but at the expense of not being able to discriminate among the errors near that area either. Finally, the issue resulting from having both A and F equal to 0 is the same as for the other two variants.

5.2.3 Accuracy Metric Functions with more Stable Denominators

5.2.3.1 Relative-Error Metrics

Fildes (1992) models the scale-dependence problem of the Squared Error (SE) function that was seen above (Section 5.2.2.1) through the following form:

$$e^2 = \epsilon^2 \upsilon \qquad \textbf{[5-5]}$$

, where ϵ are the errors due to the particular forecasting method, while υ are the errors due to the specific A recorded in the out-of-sample dataset and which can be regarded as an outlier, or as it was discussed earlier on the APE's issues (Section 5.2.2.3), a peculiarity of highly skewed datasets. Fildes makes a similar statement about how the scale-independence should be considered by saying that "… [the] *loss function* [accuracy metric] *should not depend on* υ *… any automatic forecast system should not be geared to respond to such extremes. Rather they should be dealt with by an exception monitoring scheme*" (Fildes, 1992, p.83).

Fildes analyses the SE (e^2) in the two parts as in [5-5] and then takes the square root of the geometric mean of the SE across the dataset. This is calculated by

taking the product of all SEs in the dataset, taking the geometric mean of the product and then taking its square root in order to have it in the same units as the data:

$GRMSE = \left((\prod_{t=1}^{n} e_t^2)^{\frac{1}{n}}\right)^{\frac{1}{2}} = (\prod_{t=1}^{n} e_t^2)^{\frac{1}{2n}}$, where n is the number of datapoints in the dataset.

The geometric mean is an alternative to the more widely known arithmetic mean. In the geometric mean, instead of summing all the values of a variable and dividing them by their number n, their product is taken followed by their n^{th} root.

The geometric mean has been usually applied when there is a requirement to find a meaningful average when having data that belong to different scales. In geometry, the geometric mean of two values is the square that has an area equal to a rectangle that has as its sides those two lengths of which the average is required. Eventually, since it is necessary to take the root of the product, the component values need to be positive. However, given that there are no zero values, it is possible to apply the geometric mean to either squared errors as Fildes has done with the $GRMSE$, or to absolute values as it is shown further below.

In the specific situation, Fildes, in order to make the accuracy metric scale-independent by eliminating υ, he suggests taking the ratio of the $GRMSE$ between two competing methods m and k , and get the *Relative* $GRMSE$ ($RGRMSE$):

$$RGRMSE = \frac{GRMSE_m}{GRMSE_k} = \frac{(\prod_{t=1}^{n} e_{mt}^2)^{\frac{1}{2n}}}{(\prod_{t=1}^{n} e_{kt}^2)^{\frac{1}{2n}}} = \left(\prod_{t=1}^{n} \frac{e_{mt}^2}{e_{mt}^2}\right)^{\frac{1}{2n}} = \left(\prod_{t=1}^{n} \frac{\epsilon_{mt}^2 \upsilon}{\epsilon_{mt}^2 \upsilon}\right)^{\frac{1}{2n}}$$

$$= \left(\prod_{t=1}^{n} \frac{\epsilon_{mt}^2}{\epsilon_{mt}^2}\right)^{\frac{1}{2n}}$$

The latter shows that the ratio between the $GRMSE$ metrics of two methods, apart from being unit-free, it also eliminates the peculiarities of the datasets from the comparison as they are expressed by υ. However, there are a number of necessary assumptions to adopt:

- The SE contains outliers and they are multiplicative. This is a valid assumption, since at any datapoint A , the squared error e^2 of the forecast of the location of the distribution can be expressed as the (unknown) actual squared error ϵ^2 times another positive, multiplicative value $\upsilon = {e^2}/{\epsilon^2}$, and which value has the characteristics of the datapoint
- The actual squared error ϵ^2 is assumed to be stationary while any serial relationship among the datapoints is subsumed by υ. This assumption means that there are no trends or seasonality in the dataset. In the cases examined in the present research there can be strong trends in the dataset since at the final stage of the operations the demand context might be changing to the level that the number of requests for repairs can either increase during the period, or decrease. Examples of this phenomenon have been discussed in Section 1.2 and include situations where e.g. there are probable changes in the operational demand accompanied with changes in the number of the systems supported
- The value υ affects all forecasting models equally. Even when this assumption is not fully true, the ratio between two methods eventually reduces the peculiarities of extreme values of A - at least a lot more than what was seen when the denominator was a polyonym of A like in APE, or $sAPE$

In general, an important consideration regarding the use of relative accuracy measures has to do with the practical difficulties and the potential errors when, in the presence of many forecasting model candidates, the number of ratios / pair-wise comparisons can be numerous. Alternatively, the denominator can be replaced by a benchmark model like the “random walk”/“naïve” forecast model which uses as a forecast for the next period the currently experienced value of the out-of-sample dataset. In this way, all other models are compared to the benchmark through the ratio.

Armstrong and Collopy (1992) discuss two other relative methods which use this idea of applying the benchmark model’s error as a denominator. They refer to Theil’s $U2$ measure which is a ratio of the Root of the (arithmetic) Mean Squared

Error ($RMSE$) of a forecast model to that of the naïve (the error of the naïve method is expressed as $e^* = A_t - A_{t-1}$):

$$U2 = \frac{\left(\sum_{t=1}^{n} e^2 / n\right)^{\frac{1}{2}}}{\left(\sum_{t=1}^{n} (e^*)^2 / n\right)^{\frac{1}{2}}} = \left(\frac{\sum_{t=1}^{n} e^2}{\sum_{t=1}^{n} (e^*)^2}\right)^{\frac{1}{2}}$$

Additionally, Armstrong and Collopy (1992) suggest another metric that is based on the geometric mean of the absolute errors of a model's forecast, the Geometric Mean Absolute Error ($GMAE$):

$$GMAE = \left(\prod_{t=1}^{n} |e|\right)^{\frac{1}{n}}$$

As Hyndman and Koehler (2006) and Hyndman (2006) point out, $GMAE$ is exactly the same as $GRMSE$ since in the products of squares under the squared root, the squares and the root cancel out:

$$GRMSE = \left(\left(\prod_{t=1}^{n} e_t^2\right)^{\frac{1}{n}}\right)^{\frac{1}{2}} = \left(\prod_{t=1}^{n} e_t^2\right)^{\frac{1}{2n}} = \left(\prod_{t=1}^{n} |e|\right)^{\frac{1}{n}} = GMAE$$

Now, if again the ratio is taken of the model over the naïve method as a benchmark, the Relative $GMAE$[20] ($RGMAE$) is given:

$$GMRAE = \left(\prod_{t=1}^{n} \left|\frac{e_t}{e_t^*}\right|\right)^{\frac{1}{n}} = \exp\left(\frac{1}{n}\sum_{t=1}^{n} \ln\left(\frac{e_t}{e_t^*}\right)\right)$$

Armstrong and Collopy (1992) suggest that the primary advantage of the $RGMAE$ as compared to Theil's $U2$ is the ease of interpretation, since the latter carries with it the difficulties of explaining the Squared Error (SE) as discussed in Section 5.2.2.1.

Another point that needs to be considered is in cases where the produced error is equal to zero or tends to infinity, then the whole $GRMSE$ becomes zero or infinite

[20] Even though intuitively the fact that it is relative should lead to its naming as $RGMAE$, i.e. just like (Gardner, 1990) have named the respective relative error as $RGRMSE$, the present research keeps the convention used when it was originally proposed by (Gardner, 1990), i.e. $GMRAE$

regardless of the rest of the error values. The effect of this problem can be reduced if the high and low errors are trimmed or winsorized, however, this introduces the problem of potential loss of information (Armstrong and Collopy, 1992; Hyndman and Koehler, 2006).

Davydenko and Fildes (2016) try to overcome the problem that is created by using the geometric mean for the outputs of the functions in case these are zero, by using the arithmetic mean which eventually gives the MAE. Therefore, the suggested Relative MAE for a single dataset k, is:

$$RMAE_k = \frac{MAE_k}{MAE_k^{naive}}$$

Furthermore, in order to evaluate the forecast methods using a number of different datasets $k = 1, \dots, K$, Davydenko and Fildes (2016) suggest using the geometric mean across the $RMAE_k$ and thus avoid the misleading conclusions that the arithmetic mean can bring by "averaging" non-relevant values. In order to proceed, the authors express the geometric mean of the logarithmic transformation of the $RMAE_k$ to the power of its number of data $ln(RMAE_k)^{n_k}$ - each $RMAE_k$ is raised to the power of its number of (out of sample) datapoints in order to take into consideration the difference in the weighting that needs to be implemented for each different dataset k - and calculate the average to give the Average Relative MAE measure:

$$AvgRelMAE = \exp\left(\frac{1}{\sum_{k=1}^{K} n_k} \sum_{k=1}^{K} ln(RMAE_k)^{n_k}\right)$$

Davydenko and Fildes (2016) express concerns that the measure uses the MAE and that it is out of sample data used for the denominator. Sitting behind these concerns are that, firstly if the size of the out of sample datasets is not large enough (they suggest it to be $n_k > 5$ for each $k = 1, \dots, K$), then there is a chance that the MAE of the method and that of the naïve forecast has a different level of kurtosis and thus the (natural) logarithm is a biased estimate of the ratio. In the cases used in this study it is possible to have datasets in which the horizon of the final phase can be small, e.g. it may be necessary to have forecasts for the final four months of the operations. Secondly, Davydenko and Fildes (2016) point out

the earlier discussed limitations of the MAE which uses the arithmetic mean, when the (absolute) errors' distribution is heavily skewed. In such cases they suggest trimming of the resulting outputs.

5.2.3.2 Percentage Better / Percentage Best

A different approach that proceeds in the comparison among the forecasting models, without the need of a benchmark model, is the "percentage times better" (PB). This method uses e.g. the MAE of the methods and expresses as an output the percentage number of times that each has been better than the other when compared in a number of different out of sample datasets.

It is an intuitive non-parametric measure that is also easy to calculate. Furthermore, when the data are intermittent, the measure is not affected by zeros or outliers (Syntetos and Boylan, 2005). Additionally, when there are more than two methods to compare, the percentage number of times that each method has been better than all the others are calculated, which is the "percentage times best" (PBt) measure. On the other hand, these two measures do not provide information on by how much each method is better.

5.2.3.3 Scale-Free Error Metrics

As seen above (Section 5.2.3.1), the relative-error metric approaches can have limitations if, either as a candidate forecast model, or as a benchmark model, there is a likelihood that a zero error can exist in the provided forecasts for the out-of-sample data. Even in intermittent demand data these cases are probably rare (Boylan and Syntetos, 2006), since they can still exist (Hyndman, 2006), one would want to look at alternatives.

In a number of the metrics that were studied earlier using algebraic analysis (Sections 5.2.2.3, 5.2.2.4, 5.2.2.5 and 5.2.2.6), one of the main issues that these metrics tried to solve was to reduce their dependence from the scale by using as a denominator functions of the out-of-sample values A. A different approach that the following metrics use is that they scale through the use of the in-sample data which are known in advance and the only way that they can cause problems is if they all have extreme values, for instance if all of them are zero or there is an infinite value.

The first accuracy metric of this type that is considered next is the $MAD/MEAN$ Ratio (Hoover, 2006) which uses as the denominator the in-sample mean of the data:

$$\frac{MAD}{Mean} = \frac{\frac{\sum_{t=1}^{n}|e_t|}{n}}{\frac{\sum_{j=1}^{m} in\text{-}sample\ A_j}{m}}$$

where n is the number of out-of-sample data points and m is the number of in sample points.

However, as Hyndman (2006) points out, the choice of the function to be used as a denominator should be thought of quite well. The use of the overall mean of the in-sample data assumes stationarity and thus, there is no trend or seasonality. If the data are not stationary, and there is a need to compare the forecasts several steps ahead, then the $MAD/MEAN$ Ratio metric might not be as reliable and intuitive.

Another approach has been suggested by Hyndman and Koehler (2006) which is called Mean Absolute Scaled Error ($MASE$). The authors suggest to use the MAE of the in-sample naïve forecast as a denominator:

$$q = \frac{e}{\frac{1}{m-1}\sum_{j=2}^{m}\left|in\_sample\ A_j - in\_sample\ A_{j-1}\right|}$$

and the resulting measure is $MASE = mean(|q|)$.

As compared to $MAD/MEAN$ Ratio, $MASE$ takes into consideration the peculiarities (trend or seasonality) of the in-sample data for the denominator. Furthermore, since it is usually the case that the in-sample dataset is larger than the out-of-sample, the naïve MAE is stable. Its interpretation is also relatively intuitive in the sense that a value of $MASE < 1$ suggests that the forecast of the method in the numerator gives on average smaller errors than the naïve method's in sample errors.

Davydenko and Fildes (2016) suggest that in the scenario where $MASE$ is used for the evaluation of forecasts that are produced from varying origins but constant horizon, the metric is equivalent to the weighted arithmetic mean of the

relative MAEs of the forecast method and the benchmark. This can be the case when multiple datasets (say K) are used to produce the final $MASE$ and thus their outputs need to be averaged:

$$MASE_K = \frac{1}{\sum_{k=1}^{K} n_k} \sum_{k=1}^{K} n_k \frac{MAE_k}{MAE_k^{naive}}$$

The problem then is that the arithmetic mean is overrating the accuracy and Davydenko and Fildes (2016) suggest the geometric mean as more appropriate.

Furthermore, they suggest that when the in-sample dataset is small, the value of the denominator might be small too, and thus cause the function to produce outliers.

With respect to the cases used in this study, the initial building up and the infinite-time horizon periods are long and thus they produce enough in-sample data. Furthermore, any accuracy metric used would be applied to evaluate forecasts for different types of spares and different ranges of values. Therefore, a scale-free type of metric was required. MASE was chosen as a good candidate. This is because apart from being scale free, it also does not require many pairwise model evaluations like the other scale-free metrics do like the Relative-Error (Section 5.2.3.1), while it gives a more precise value of how much better one model is compared to another, something that would be missing if the Percentage Better / Percentage Best metrics were used (Section 5.2.3.2). Consequently, the geometric mean of the $MASE$ outputs of the final phase out of sample / test data has been used for evaluation.

5.2.4 Accuracy Implication Metrics

5.2.5 Accuracy Implication Metrics Using Supply-Provision Objectives

The following observation by Gardner (1990, p.492) that "… *forecast errors are the primary determinant of the safety component investment. In general, the better the forecast accuracy, the smaller the inventory investment needed to reach any particular target value for customer service*" stands out from others in that it places emphasis on commercial realities such as that the forecast models

are rarely an end on their own, but they usually are a means to an end, that of facilitating decision making in the provision of logistics support.

There are three interrelated points in Gardner's observation. Firstly, it is the "*targeted value for customer service*". This is the reason for which inventories are operated when used for support. The second point is a related competing objective, i.e. that the inventory investment/costs need to be small. Finally, that the forecast errors are a very influential factor on how much safety stock is maintained in the inventory.

This final point deserves further consideration. The planned safety stock is defined by the assumed probability distribution demand model and the set type of service metric along with the set service level. The latter is the decision resulting from the trade-off between the customer service that is desired to be offered, and the related costs of offering this service. These relationships are summarised in Syntetos, Nikolopoulos and Boylan (2010, fig.1): the service level and the costs are determined by the interaction of the (accuracy of the) forecast method and the inventory rules through the inventory's stock management system. However, as shown below (Sections 5.2.5.1, 5.2.6.1), the accuracy implication metrics that have been commonly applied in the literature are not enough for the FPP cases to operationalise all the dimensions of the problem related to the quality of the forecast, and additional factors need to be included in the assessment. In the following paragraphs, the principles that Gardner (1990) describes, are compared to the FPP cases to determine where and how they differ.

5.2.5.1 Final Phase Idiosyncrasies 1

Gardner (1990) developed/demonstrated the idea of using the two objectives of the inventory investment and the provision of customer service to evaluate the provided models' forecasts, by using the total costs TC over a set time period as presented in **Hadley and Whitin (1963, chapters. 2, 4)**.

$$TC = A\left(\frac{\lambda}{Q}\right) + IC\left(\frac{Q}{2} + R - \mu\right) + \frac{\pi\lambda}{Q}\left(\int_{R}^{+\infty} xd(x)dx - RD(R)\right) \quad \textbf{[5-6]}$$

While Gardner (1990) uses the delay time as a measure of customer service, other researchers have used different service levels, like the backlog volume (Kourentzes, 2013) etc. Nevertheless, the idea is to have the customer service measures expressed in such a way that lower values signify better results. In this way, the measure of service used can have the same direction of better quality as its twin objective that indicates the costs of having inventory and thus both be able to be used in a plot like the one in Figure 5-5.

Equation [5-6] calculates the average total costs of keeping inventory for a single item/component per unit time (say quarter) and in infinite-time-horizon settings. The variables involved are:

A are the re-order costs, i.e. the costs paid every time a new order is placed

λ is the average demand for the item during the period / unit time that the inventory is used

Q is the amount ordered for this item every time an order is placed

I is the inventory holding cost rate. It includes storage costs (warehousing rents, insurance, etc.), obsolescence and tied-up capital as an opportunity cost

C is the purchase cost of every unit of the item

R is the reorder point. It is estimated by taking into consideration the planned service level for the item and the item's demand during the effective lead time (which, in the problem setting examined by Gardner (1990), the effective lead time is equal to the lead time **(for a discussion on the relationship between the lead time and the effective lead time see e.g. Waters (2011, pt. 5)**)

μ is the expected (mean) demand during the lead time, i.e. the number of units that are expected to be consumed during the lead time

x is the random variable of the realised demand during the lead time, i.e. the number of units that can be consumed during the lead time

$d(x)$ is the marginal distribution of the lead time demand x

$D(x)$ is the complementary of the cumulative distribution of the lead time demand $d(x)$, i.e. $D(x) = 1 - \int_{-\infty}^{x} d(x)dx$

π is the backordered cost/"penalty" per unit item backordered

> Hadley and Whitin (1963, p.19) suggest that a general function for the backordered costs should depend both on each unit and on the length of time during which the backorder is active: $\pi(t) = \pi + \hat{\pi}t$, with $\hat{\pi}$ being the rate of increase of the cost per unit time that the backorder is not covered. However, for the estimation of [5-6] the author assumes that the backorder cost depends only on the number of units and not on the duration that the backorder is still on. Later on it is shown that the time dependence needs to be considered in the FPP type of problems

A brief discussion about the terms in [5-6] now follows in order to assist in identifying any possible differences that the FPP cases of interest might have.

The first term expresses the (costs from the) number of orders that can be placed in the set period of time. In more details, in the examined period, orders of size Q are placed whenever the inventory falls to or below the reorder level R. So, if λ is the average demand for the whole period, the number of orders is $\frac{\lambda}{Q}$. However, this term is not affected by the forecast accuracy because the forecast is used for the duration of the effective lead time and not for the period that the inventory is used. The effective lead time is the lead time and – in case the inventory policy is a periodic review – it is also the review time.

It is the other two terms that depend on the forecast and the smaller they are the better. The second term is the holding cost and is associated with the first of the two in the twin objective, i.e. the inventory investment, which is a measure of the efficiency of the supply system. The third term has to do with the penalties incurred from unmet orders. In essence it is a proxy of the (reverse of) the customer service quality, the effectiveness of the supply system. The fewer unmet order penalties incurred the better the service provided by the inventory. What Gardner (1990) showed and has since been repeatedly applied (Eaves and

Kingsman, 2004; Kourentzes, 2013; Sani and Kingsman, 1997; Syntetos, Nikolopoulos and Boylan, 2010; Syntetos and Boylan, 2005), is that different forecast model accuracies, for the same systems, give more/less accurate forecasts and thus less/more inventory and back-order volumes.

An example from Boutselis and McNaught (2018) is replicated in Figure 5-5. The forecasting model whose resulting curve is closer to the axis is better in the accuracy implication metrics.

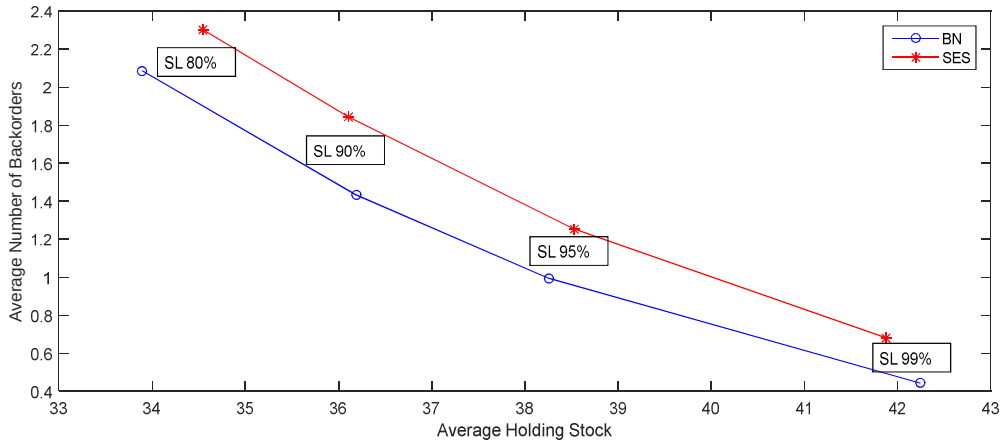


Figure 5-5: Inventory investment (holding stock) vs number of backorders under different target service levels (SL) of two forecast models

Actually, it is the infinite-time-horizon nature of the problem that renders the first term $A\left(\frac{\lambda}{Q}\right)$ uninfluential in the study about the possible forecast accuracy implications. Additionally, in **Hadley and Whitin (1963, chap. 2)** there was an extra additive term in the total costs equation [5-6], the λC, which was used in order to include the costs of the item's units purchase cost for the period that the inventory is used. This term was not further considered by Hadley and Whitin, since it would not affect the inventory holding decision policies on the number of items to order Q and on the level of the re-order point R. In other words, in an infinite-time-horizon λ is going to be consumed anyway – the time is infinite - so it does not affect the comparison among the inventory holding policies, neither it is directly related to the forecast of interest, which is referring to the effective lead time.

On the other hand, in the FPP cases, the situation is not the same: the time-horizon is finite. In essence, for the FPPs the whole period is an effective lead time and this imposes a number of non-relaxing restrictions that increase the decision making risk, like the inability to rely on a backorder that can soon be covered, or that there is no subsequent period to consume any likely leftovers from this period. Consequently, as it is shown further below, the forecast accuracy does not only affect the service level (as approximated by the backorders' volume) and the holding costs (as approximated by the inventory volume held). It also affects - intuitively as well - the number of units purchased at the beginning of the period, i.e. the respective to the λC that was excluded from the consideration of the costs in [5-6], and thus, the number of unused leftovers at the end of the period or the number of stock-outs occurred.

In order to show the dependency of the decision on additional factors to what the steady-state situations of the infinite-time-horizon do, once more reference is made to **Hadley and Whitin's (1963, chap. 6)** work on "Single Period Models" that are also well-known as the Newsvendor Problems (NVPs).

So, to develop an argument for the additional dimensions that need to be considered, it is necessary to view the NVP from the generic seller's perspective. The gain function version of the NVP problem is being used here in [5-7], as opposed to the costs function version of [5-6], so as to use the "loss of good-will" cost π (equivalent to backorder costs of [5-6]) explicitly, instead of the opportunity costs that are considered of the equivalent NVP problem formulation as a cost function (see e.g. Lodree, Kim, and Jang (2008), or Khouja (1999)).

To present the profit/gain function $G(Q)$, of the Newsvendor decision maker, it is necessary to introduce some additional notation:

S is the selling price for each unit of the item under consideration. For the cases of the "final phase" that are of interest, it could be considered as some notional operationalisation of providing a unit spare when it is needed

L is the price that any items left at the end of the period are sold, with $L < C$. Again, for the cases of interest to this study, this factor could be considered as a notional use of an item after the end of the period of interest. However, if such a

use does not take place, its value could be set to zero, or even be negative, if there is some kind of disposal cost, or cost of having to pack the unused stock and transport it to another place after the operations are over

Using this notation and the notation of [5-6], it is possible to write the gain function $G(Q)$ for any amount Q that the Newsvendor might decide to purchase:

$$G(Q) = S\sum_{x=0}^{Q} xd(x) + SQ\sum_{x=Q+1}^{+\infty} d(x) + L\sum_{x=0}^{Q-1}(Q-x)d(x) - \pi\sum_{x=Q}^{+\infty}(x-Q)d(x) - CQ \quad \textbf{[5-7]}$$

None of the above terms depends on time as the inventory holding costs do in [5-6]. Regarding the first two terms in [5-7] the first has to do with the gain if the demand during the period is less or equal to Q, while the second term expresses the gain if the demand is more than Q. A thing to observe is that in contrast to [5-6], the [5-7] is a gain function, and these two terms are an additional consideration due to requirement to include what the Newsvendor can "sell", and which does not depend to the difference in the type of time-horizon between [5-6] and [5-7].

The third term has to do with something which is a peculiarity of the NVPs and thus, an additional concern for the FPP cases as well. This term expresses the leftovers that might remain after the end of the period. Consequently, a better forecasting model, under the same inventory policies should help produce fewer leftovers after the end of the period than the rest of the models. Furthermore, the weight that the modeller should give to such an event should depend on if the value of L could be considered positive, zero or negative.

Finally, as in [5-6], the fourth term is similar but not exactly the same as in the infinite-time-horizon cases. In the infinite-time-horizon this would be the backorders placed, but in the finite time-horizon there is no subsequent period to fulfil any not covered demands. In these cases, they represent the "loss of good

will"/stock-out cost which can be considered equivalent to the "lack of service" approximation through the volume of backorders. Nevertheless, for the present research use, this term has been treated the same as the backorders' volume used in other cases in the literature.

However, as **Hadley and Whitin (1963, secs. 6-5)** explain, the NVP formulation can be better estimated by the inclusion in [5-7] of the following time-dependent costs (TDC) as well:

$$TDC(Q) = ICT\left[Q - \frac{\mu T}{2}\right] + (IC + \hat{\pi})B(Q) \qquad \textbf{[5-8]}$$

, where $B(Q)$ is the probability of a stock-out, given different values of Q, and $\hat{\pi}B(Q)$ is a time-dependent stock-out cost.

It should be emphasised, and as it has also been mentioned in Chapter 2, that due to the inhibiting distances from the resupply centres of the Support Chain and the fact that the operations are about to finish and thus, the decision makers would usually restrain from placing additional orders. Consequently, the time-dependent stock-out cost rate $\hat{\pi}$ of the second term in [5-8] mainly expresses things like the rate at which operational outcomes are decreased when a required spare is not available to make the system available for use.

Finally, the first term of [5-8] expresses the holding costs in a similar way as in the infinite-time-horizon cases.

As seen in the infinite-time-horizon cases of [5-6] there is a dual-objective of interest that the quality of forecasts from the different models can affect, namely the backorder costs that reflect the service level (expressing the effectiveness) and the average inventory holding costs (expressing the efficiency). In these cases, due to the lack of knowledge of the cost/monetary factors, these two objectives (backorder costs and holding costs) are approximated by the volume of inventory backordered and held respectively during a certain typical period of the infinite horizon.

Now, regarding the FPP cases, these two objectives (the volume of inventory backordered and volume of inventory held) are still important as seen by the

discussions on [5-7] and [5-8], with the only difference that the average backordered volume is described as average stock-out volume. However, as shown there are two additional factors that need to be included as implications of the forecasting accuracy. Firstly, it is the number of items which are "left" at the end of the period under consideration. This is an economical/efficiency measure and in an ideal situation would have zero such items, so, under the same conditions, the higher it is the worse the implication of the forecast model accuracy. Secondly, it is the time dependent stock-out effect. This should reflect the duration that a stock-out problem has existed for, which is related[21] to what Schneider (1981) calls α-service level and Axsater (2006) calls S_1 service level, i.e. the probability of no stock-outs per cycle. Consequently, this measure reflects the lack of service from the inventory to the mechanic that needs a spare and does not find it. A point that needs to be highlighted is that in the FPP cases, duration that a stock-out problem has existed for is related but is not exactly the same as the probability of no stock-outs per cycle. This is because when there is an infinite-time-horizon, the event of having no stocks on the shelves at any period is a random variable, while in the finite time-horizon of the FPPs, the last period is more likely than the one before the last etc. Nevertheless, in the present research the number of periods that have zero inventory on the shelves are calculated to approximate this specific accuracy implication.

In summary, in the FPP cases a forecast model's accuracy implications could be evaluated using the following four measures:

1. The average volume of stock-outs during the FPP period, representing the number of cases the mechanic did not find the part "on the shelf" (a measure of the effectiveness)
2. The average volume of inventory during the FPP period, representing the costs of keeping inventory (a measure of the efficiency)

[21] They are related but not exactly the same because when there is an infinite-time-horizon, the event of having no stocks on the shelves at any period is a random variable, while in the finite time-horizon, the last period is more likely than the one before the last etc. Nevertheless, in the present research the number of periods that have zero inventory on the shelves are calculated to approximate this specific accuracy implication

3. The volume of spares left at the end of the FPP period, representing the amount of unneeded inventory (a measure of the efficiency)
4. The probability of no stock-outs during the FPP period, representing the intensity of the problem to the mechanic caused by the inventory low service (a measure of the effectiveness)

Nevertheless, in order to summarise the evaluation it is necessary to average each of the above four metrics over all the different components. On the other hand, this simple overall averaging does not accurately reflect the effects on the support of the operations, since the components can be highly interdependent, in the sense that e.g. stock-outs on a component can make a supported system not available and thus reduce the demands for the other components.

Furthermore, the overall average – either geometric or arithmetic – might not represent the real service level provided by the inventory either. For example, low volumes of backorder in one or more components might not be a result of the good synergy of forecasting methods and inventory rules within the stock management system, but it might be due to the fact that another component occasionally runs out and keeps many of the operated systems unavailable. Consequently, an overall average of the volume of backorders would be biased since it would possibly give a low value and thus a good overall service. On the other hand, such a case would not necessarily result in high volumes of inventory in order to indicate a problematic situation.

As shown in the following section, there are types of Support Chain relationships, like the ones of interest to this research and discussed in Sections 1.2, 1.4 and 3.4, in which the interdependence consideration in the forecasting evaluation can be included. In such cases the service level measure 1 above is considerably simplified from averaging for all the items into incorporating them into the Operational Availability of the supported systems.

5.2.6 Accuracy Implication Metrics Using the Systems' Operational Availability as Objective

The importance of the systems' Operational Availability objective as a measure for the evaluation of the forecast models' accuracy has been used by a number

of practitioners (see e.g Systecon (2015), Sherbrooke (2004)) and it reveals an important difference in the point of view between what was discussed in the previous section and here.

Sherbrooke (2004) suggests that the supported systems' (aircraft in Sherbrooke's case) Operational Availability should be used as an objective measure for the demand forecast evaluation on a set of components, since the ultimate objective of the whole support operations (logistics and repair related activities) is to keep as many of the supported systems ready and operational as possible using a given budget. Therefore, the Operational Availability is a measure of the SC effectiveness.

The difference found between the measures discussed in Section 5.2.5 which use the individual components and what Sherbrooke suggests about the supported systems, can be attributed to differences in interests of the decision makers. In the first instance of Section 5.2.5, the decision maker is interested in meeting the service level requirements of a supply provision contract in an efficient way, and thus the interest lies in the discussed trade-off of individual components' (lack-of) service level/stock-out volume (effectiveness) and the inventory costs/inventory volume (efficiency). On the other hand, when the decision maker's focus is in the whole Support Chain up to and including the Operations - like the cases that Sherbrooke (2004) refers to, or in "availability" type of contracts - then the interest lies in the different levels of systems' attained availability (effectiveness) that the forecast models can help realise, given the same investment in spare components (efficiency).

This observation highlights both the difference in the decision makers' interests but also signifies the difference in the forecast accuracy implication measures that each decision maker view values more. As mentioned in Chapter 1, the research problem's interest is in the "final phase" of operations which are supported in logistics and repair, and thus, from that perspective the present research is more related to Sherbrooke's view.

Furthermore, there is an additional advantage/convenience to be gained from using the measuring approach that Sherbrooke suggests. In certain cases, e.g.

when the decision maker serves a contract that does not include the consideration and the related ability to have access to data regarding the Operational Availability of the supported systems, the two objectives of inventory and of backlogged volume, plus 3 and 4 of the list presented earlier in case there are NVP and FPP types of problems, can be the only choices. In such cases, each forecast model needs to be evaluated through all these objectives by averaging over all spares/components examined. On the other hand, in the cases where data on the Operational Availability of the systems are accessible by the decision maker, the main output of interest is only a single “thing” – the system that the logistic and repair activities support, so that it stays available and operational. It is this system that “amasses” in a natural way all the components’ inventory contribution in one. This conveniently removes any ambiguity that might exist when averaging over different units of interest which are thought dependent.

5.2.6.1 Final Phase Idiosyncrasies 2

Nevertheless, when compared to the infinite-time-horizon that Sherbrooke examined, there are still some differences in the “final phase” cases that were considered in the present research. Actually, these considerations have been included in related Multi-Indenture Multi-Echelon (MIME) optimisation models like the “Endurance” model (Systecon 2015). In such models there are two optimisation objectives of a single finite time-period of interest related to the effectiveness of the SC:

1. The deployed/supported systems’ Operational Availability at the very end of the period
2. The deployed/supported system’s average Operational Availability for the duration of the period of interest

This idiosyncrasy of the FPPs as compared to the infinite-time-horizon problems that Sherbrooke (2004) was referring to, can be seen as a special case of the related discussion in Boylan and Syntetos (2006) and Willemain (2006) on whether to forecast for the mean and the variance of the demand distribution, or for certain extreme percentiles. In the case of the present research, it is the high percentiles accuracy implications of the demand forecasting models that need to

be observed. This means that our interest is in the problems incurred by low values of Operational Availability.

So, what is suggested is to compare the candidate forecast models under the above two objectives, replacing number 1 in the list of Section 5.2.5.1. In summary, in the present research for the FPP, and under the assumptions of close relationships within the SC (as referred to in Sections 1.1, 1.2 and 3.4), the following accuracy implication metrics have been used:

1. The average volume of inventory during the period, representing the costs of keeping inventory (a measure of the efficiency)
2. The volume of spares left at the end of the period, representing the amount of unneeded inventory (a measure of the efficiency)
3. The probability of no stock-outs during the period, representing the intensity of the problem to the mechanic caused by the low service (a measure of the effectiveness)
4. The systems' Operational Availability at the very end of the period, representing the intensity of the problem to the end-customer (a measure of the effectiveness)
5. The systems' average Operational Availability for the period of interest, representing the intensity of the problem to the end-customer (a measure of the effectiveness)

Finally, in order to create pairwise plots as a useful presentation tool to facilitate decision making, measures 1 and 5, and 2 and 4 were paired. The 3rd measure was used as a complement to the two pairs.

5.3 Conclusions

Chapter 5 considered two ways of evaluating forecasts, namely with accuracy metrics and accuracy implication metrics.

Regarding the accuracy metrics, it was shown that by doing an algebraic analysis of their loss functions, the analyst can understand the areas where they can be applied, but most of all their limitations. Such intuition was provided by showing that the $sAPE$ loss function, in any of its variants should not be used for error

magnitudes larger than the value of the data point that is to be forecast, while all three variants are indeed identical to the rest of the errors' value space.

Furthermore, regarding the chosen accuracy metric, it was suggested that since the datasets that are used in the FPP are outputs of the demand for not just a single type of spare part, one would expect different magnitudes of demand, and therefore, the chosen accuracy metrics should be able to accommodate such a requirement. In addition, the simulation used in this thesis produced datasets from the multiple runs of every simulated future scenario (see Section 1.3). Consequently, there were different sets of time-series, and this was an additional challenge to the comparison of the forecast models' outputs that the accuracy metric should be able to accommodate as well. Therefore, scale-free accuracy metrics were required, and $MASE$ (Section 5.2.3.3) was the one that was chosen, because, as compared to Percentage Better / Percentage Best metrics (Section 5.2.3.1), it is able to give the magnitude of difference among the candidate forecast models' outputs, while it does not need to do multiple pair-wise comparisons as the Relative-Error metrics do (Section 5.2.3.2).

It was also shown that the FPP requires a number of accuracy implication metrics in addition to the existing indicators in order to evaluate the quality of a forecast. These are the volume of spares left at the end of the period, the probability of no stock-outs during the period, and the systems' Operational Availability at the very end of the period.

6 SIMULATION

6.1 Introduction

Chapter 6 presents the Support Chain (SC) system that is simulated in order to generate the data required to build and to evaluate the forecast models. The chapter also describes the Activity Diagram that was used to build the simulation as a computer model.

6.2 Simulated System

Using the literature review from Chapter 2 and the SMEs' interviews from Chapter 3,

a number of factors that could be influential in the realisation of the demand (Table 2-1 and Table 3-1) were identified, as well as three conceptual models. The intention has been to study how these factors work in a Support Chain (SC).

In order to proceed, the option was taken to collect data from a simulation that was developed of such an SC. The use of simulation data is obviously less realistic than using real-life data. Real data would have the advantage of increased credibility of the results, particularly in the eyes of practitioners.

On the other hand, in order to compare forecasting models, a simulation offers a number of benefits as compared to real data. Firstly, real data can include a number of errors and anomalies that can be difficult to identify as such, and these can cause misjudgements especially when the research is in areas with very little background experience. On the other hand, through the use of a verified simulation, the chances of errors are reduced. Moreover, in case of an unexpected and debatable outcome, the ambiguity can always be resolved by the simulation's capability to replicate the runs, or even investigate them step-by-step. Such a benefit is very important. Low noise in the data increases the credibility of the evaluations in the forecasts, the results and the conclusions.

Furthermore, real life data would restrict the evaluation of a study to a single realisation of the Support Chain system as well as the ability to investigate different system's settings of interest, or, as in the current research, different possible final phases. Having the convenience to use a tool like a simulation, the study can be upgraded into an experiment. Different settings of the system can be examined and each can be replicated a number of times in order to explore the aleatory uncertainty. This

capability benefits the research in a number of ways. Firstly, the range of situations researched is expanded and thus our acquired understanding increases as well; a number of insights resulting from the simulation experiment are presented in Appendix B. Secondly, there is a wider range of situations in which to compare the forecast models and this increases the power of the statistical tests when investigating differences between these models.

Additionally, even though the development of a simulation model can be an elaborate task, it may still be less demanding in time and resources compared to what might be needed in order to collect and process the real-life data from the logbooks. Real-life data would require access to multiple nodes within the SC and the acquisition of the data from the logbooks may not be in electronic format to process. Furthermore, there would be a need to cleanse and synchronise the data. Finally, real-life data can also be of a sensitive nature, and therefore accessibility might be restricted or even denied.

The simulation model (Section 6.4) was built for two uses. Firstly, it provides a means by which to generate the data required to build the demand forecast models. These were the in-sample data as briefly discussed in Section 5.2.1. This involves a scenario being simulated just once to generate data corresponding to what might be collected in logbooks in the various phases of operations before the final phase.

Secondly, as suggested above, the simulation offers the ability to examine different possible futures in the form of an experiment. The resulting generated data provides the out-of-sample data, again as discussed in Section 5.2.1. Therefore, the simulation model was also used to generate data corresponding to several different possible final phases by changing factors that could realistically have been changed as a result of the operations coming to an end. Such factors were the number of operational systems, the number of mechanics, the level of spares in the inventory, etc.; these are listed in more detail later. As discussed in Chapters 1 and 2, the operational demand does not always decline in the final phase. Consequently, the operational demand was included in the set of factors that were examined. Finally, in order to allow for the aleatory uncertainty associated with the final phase, it was necessary to run each possible final phase multiple times, i.e. 100 replications.

The following section (Section 6.3), describes the factors that were taken into consideration in order to develop the simulation model, along with their assumed range

of values. As presented in the same section, the scenario involved the support of a number of UAVs that were performing ISTAR operations. The values used do not come from real systems. However, they correspond to small-scale operations in which a typical 4-hours ISTAR mission per UAV could be used twice a day to provide 24-hours' coverage. Consequently, three to four UAVs were considered with a corresponding level of support.

Section 6.4 presents the Activity Diagram that was used as a basis for the simulation. The scenarios experiments design along with the results and their analysis are then discussed in Sections 7.2 and 7.3.

6.3 Factors Included in the Simulated Scenarios

One of the conceptual models referred to in Section 3.3 and which is pictorially presented in Figure 6-1, was used in order to identify the factors to include in the simulation model.

This conceptual model considers the four different contexts within which the supported systems exist. These contexts are presented schematically in Figure 6-1

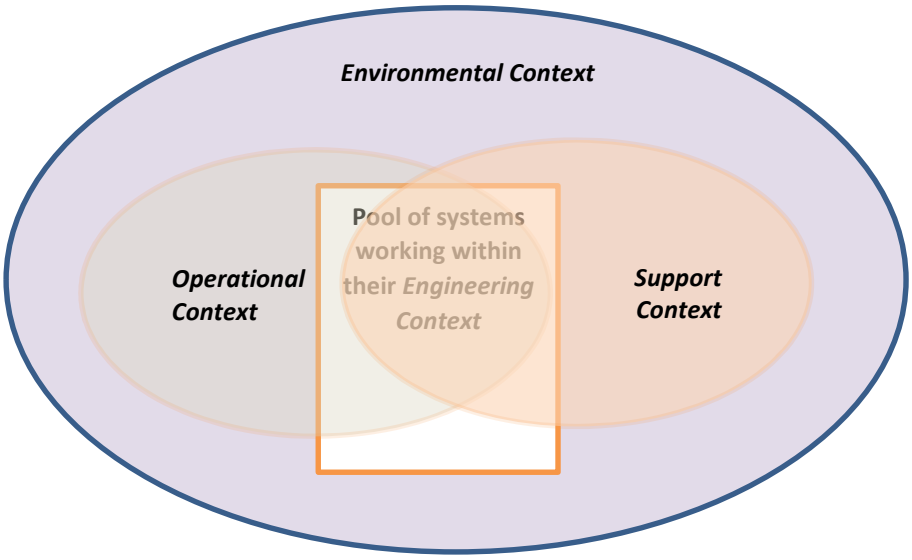


Figure 6-1: Conceptual model of the sources of demand context

In the Sections that follow, each context is examined in order to present the factors that were included in the simulation model.

6.3.1 Engineering System Context

The first is the engineering system context to which the supported systems belong. From this context it is possible to identify the factors that are related to the Equipment Breakdown Structure (EBS) of each system. Consequently, what was included in the simulation model were the components' reliability and maintainability (SN 6 Table 2-1).

In order both to include a variety of component types, but also to make the model more manageable, it was assumed that the simulated operation had a number of fictitious Unmanned Air Vehicles (UAV)s and that each one was composed of the following components:

- One Line Replaceable Unit (LRU). Such items are fundamental for the operation of any system in the sense that if they do not work, then the system – the UAV in the present case - cannot operate. Consequently, in a repair activity concerned with an LRU, the UAV is repaired by replacement with another LRU from the stock of spares. LRUs are also repairable. It has therefore been assumed that in order to repair an LRU, one Disposable Part (DP) was required. It was also assumed that the LRU's time-to-failure followed an exponential distribution with a Mean Time Between Failures ($MTBF$) of 80 hours. Moreover, it was considered that the components did not age for the period of interest. Additionally, a diagnostic activity was assumed to precede the repair activity. This was assumed to be always correct and to have a duration following a truncated normal distribution with a mean of 1.2 hours, a standard deviation of 0.7 hours and a minimum value of 0.2 hours $N(1.2, 0.7^2, min = 0.2)$. Finally, in order to replace the LRU on the UAV, the activity duration was sampled from $N(2, 0.8^2, min = 0.3)$

  As stated previously, the LRU was assumed to be composed of a single DP. Obviously, since this part would be the only reason for the LRU to malfunction, it had the same reliability characteristics. The duration for diagnosing that the DP would have been the cause of the LRU's fault was sampled from $N(2.5, 1^2, min = 0.4)$ and the duration to replace it was sampled from $N(2.5, 0.9^2, min = 0.3)$

- One Partly Repairable Unit (PRU). Such items are also fundamental for the operation of any system, and in a repair activity the UAV is repaired by

replacement from the stock of spares. PRUs are repairable as well, but they do not need another part for their repair. An example of such components can be those that need a careful calibration at the repair shop, for instance a compact gyroscope.

It was assumed that the time-to-fail of a PRU followed an exponential distribution with a Mean Time Between Failures ($MTBF$) of 100 hours. It was further assumed that the PRUs did not age for the period of interest, and that the diagnosis' duration followed $N(2.8, 0.5^2, min = 0.3)$. Additionally, in order to replace the PRU on the UAV, the duration was sampled from $N(3, 0.6^2, min = 0.4)$.

As mentioned above, PRUs are repairable. Diagnosis of the problem was assumed to take a duration of $N(3.5, 0.8^2, min = 0.45)$, while repair time was sampled from $N(4, 0.7^2, min = 0.45)$. However, PRUs are not always repairable. It was assumed that the probability of a PRU being repairable was 0.8

- Finally, it was assumed that the UAVs also contained a single Disposable Unit (DU), e.g. landing gear/tyres, which are also fundamental for a UAV's operation. Such components are not repaired after they are replaced by a spare on the UAV. It was further assumed that the time-to-fail of a DU also follows an exponential distribution with a Mean Time Between Failures ($MTBF$) of 110 hours. Again, assuming that the DUs do not age for the period of interest. Furthermore, it was assumed that the duration of diagnosis follows $N(1.7, 0.1^2, min = 0.2)$. Additionally, in order to replace the DU on the UAV, the duration was sampled from $N(1.5, 0.2^2, min = 0.2)$

The simulated scenario assumed that each UAV could fly a continuous four-hour mission (SN 4 Table 2-1).

The simulation scenario assumed that the following recorded incidents would be found in the "logbooks" of the nodes of the Support Operation:

- Fault of the UAVs due to LRU, PRU, DU, and from which the respective failure rates could be acquired from the outputs of the simulation (for the notation see

Table 6-1): FRT_LRU, FRT_PRU, FRT_DU and also the number of hours that the components had operated until they failed: $FlHbd\_LRU, FlHbd\_PRU, FlHbd\_DU$.

- Fault of the LRU (due to the DP), fault of the PRU: $WFRT\_LRU$, $WFRT\_PRU$
- The decision to discard the PRU when it is beyond repair: PRU_disc.

6.3.2 Operational Context

The second context is the operational one (Figure 6-1). The operational context is the one that uses and "wears-out" the components of those UAV systems that are available. Three relevant factors that are related to this context were identified (SN 4, 5 Table 2-1). Firstly, the number of systems (xNU) and the number of operators (xNP) that are deployed. It was assumed that these numbers vary during the scenario and their values are provided when subsequently described. However, it was always assumed that there were as many operators as there are UAVs. The second factor is the operational demand ($OpDem$). It was assumed that the UAVs were deployed in support of ISTAR (Intelligence, Surveillance, Target Acquisition, and Reconnaissance) operations and the requirement was expressed with the coverage of a certain area from the air for a specified percentage of 24 hours every day. Again, the scenario's $OpDem$ levels are subsequently described in more detail. Finally, the information provided from the interviews (SN 8, 9 Table 3-1), which showed that the experience of the operators was a factor that contributes to the spares' demand ($PExp$), was also included. It was assumed that there were three levels of experience, 1, 2 and 3, with a probability of having each distributed as 0.2, 0.43 and 0.37 respectively. Each level would have a different effect on the wear-out of the system, with the most experienced level 1 operator having no degrading effect when he/she operated the UAV, level 2 having a 20% effect and level 3 a 40% effect.

In the respective logbooks, it was assumed that the following incidents and records could be acquired:

- The operational demand: $OpDem$
- The take-off and landing events (either due to normal, or to emergency landing), from which the operational rate and the duration of each flight / Time on Task could be acquired: $OpRT, TOT$
- The operator that controlled the flight, whose level of experience could be operationalised from his or her rank: $PExp$

- The number of units deployed: xNU

6.3.3 Support Context

The third context (Figure 6-1) is that of support in which the maintenance and logistics/supply activities take place and keep the systems available. This was assumed to be the largest of the four contexts because it was expected to include both the repair and the supply/logistics policies, activities and resources (SN 11-15 Table 2-1).

The repair functions were driven by the EBS of the system. The repair policy is one of the factors that was shown in Chapter 2 and has been identified as influential, but it was assumed not to take changes into consideration. Therefore, included was a single, unchanging policy into which there was a first-line repair through replacement of the components (LRU, PRU or DU) and it was assumed that the reason for a fault can only be a single one and was always identified correctly.

However, a simulation was carried out for two different levels of experience of the mechanics ($MExp$) that perform the diagnosis of the faults and the repair; level 1 and level 2 in a 55-45% split, respectively when the number of mechanics changed during the scenario (SN 1 Table 3-1). Level 1 mechanics had no increasing effect on the duration of the diagnosis, while the less competent level 2 mechanics had a 40% increase in the mean duration both of any diagnosis or repair that they did. The relevant durations were the ones described earlier in the engineering context (Section 7.3.1). Apart from experience, it was assumed necessary to consider the resources, which in this case included only the number of mechanics (xNM). The pool of mechanics deployed was changed as the steady-state phase of the scenario evolved, and details of that change follow.

Since there were components that are repairable (LRU and PRU), consideration was also given to a second-line of repair of the components. In the second line, there were repair processes and resources. The processes in the simulation were not varied, so, it was assumed that there was a priority to the first line, but, when a mechanic took over a job he/she followed it till the end. This enforced a “no-batching” repair policy, similar to the “no-batching” supply policy that is now described.

The supply/logistics functions were composed of two lines. One line was right next to the second-line repair shop – a second-line depot – in which were held the inventories

of the four spare components: LRU, PRU, DU and DP. The no-batching policy was the common (S, S-1) policy in which whenever there is a single reduction in the level of the inventory position[22], a resupply (or repair) order was issued. The target level of inventory was another factor (resource) that was considered as the scenario evolved ($xSLRU, xSPRU, xSDU, xSDP$).

The resupply was coming from a third-line which was assumed to have unlimited resources. The "logistics" connection between the front-line and the second-line took place through a limited number of transport vehicles which was another factor that was taken into consideration ($xNTr$). The duration of the one-way transport was sampled from $N(2, 0.05^2, min = 0.8, max = 12)$. Furthermore, it was assumed that it was the driver that also searched for the spare part in the depot and it took her/him $N(1, 0.6^2, min = 0, max = 4)$ hours to find it and load it on the vehicle. Finally, the connection with the third-line was performed with unlimited resources but with a duration that followed[23] $N(168, 24^2, min = 30)$.

The related records from the logbooks were inevitably more:

- As referred to previously, the activities from the engineering context were FRT_LRU, FRT_PRU, FRT_DU, $WFRT\_LRU$, $WFRT\_PRU$, PRU_disc. But from the logbooks it was also possible to get the following repair durations: Rdu_LRU, Rdu_PRU, Rdu_DU for the repair of the UAVs at the first-line and $WRdu\_LRU, WRdu\_PRU$ for the durations of repair of parts at the second-line.
- The repair resources were also considered as available next to the incidents. The data availability also included the experience of each mechanic that took over a job: $MExpB$ and $MExpW$ for the first-line and second line respectively, their total numbers xNM and the condition of the mechanics' pools during each record: $QM, BWkld, WWkld$, which were the number of mechanics who were idle, the percentage that worked in the first-line and the percentage that worked in the second-line.

[22] The inventory position is the on-hand inventory plus the ordered but not yet arrived amount (the "due-in") minus the backlogged

[23] The 168 hours are equal to a whole week and it is the duration from the time the resupply order is placed until it arrives on the shelf of the depot

- Similar records were for the logistics' administrative resources: $xNTr, QAdm$, which were respectively the number of deployed vehicles and the number that were idle at any specific time.
- Regarding the inventory, interest was in the levels: $xSLRU, xSPRU, xSDU, xSDP$ and the respective on-hand units: $Oh\_LRU, Oh\_PRU, Oh\_DU, Oh\_DP$.
- Furthermore, there were spare orders as incidents and these could result in the estimation of rates as: ORT_PRU, ORT_DU and ORT_DP, for order rates for PRU, DU and DP respectively.

6.3.4 Environmental Context

The final context (Figure 6-1) that was examined was the environmental one. As discussed in Chapters 2 and 3, the environment does not only affect the life of the components (i.e. what is included in the engineering context), but as discovered in the interviews with the SMEs, it also has effects upon the other two contexts.

Two levels for the environmental conditions were considered in more detail, level 1 and level 2, with the latter assumed to be worse. Therefore, whenever a UAV was operating, apart from the effects of the operator's level of skill, it also included a 30% further increase in the degradation of all its components if the environment was level 2. Additionally, if any kind of transport was taking place under these conditions, a further 20% increase in the duration was included. The percentage of time that each level of environmental conditions took place was 60% and 40% for levels 1 and 2, respectively, while the level was assumed not to change during a whole day. A record that was available in this case was the level-condition of the environment which was called: Env.

A list of the factors that have been included in the simulation and which possibly affect the demand for spares is presented in Table 6-1. These are possible factors and not certain in the sense that in the scenario in which they interact they might not appear to be very influential, and identifying which of them are, is one of the benefits of building and investigating using models (e.g. see Section B.1).

Table 6-1: Nomenclature

$\underline{\boldsymbol{OpRT}}$: Operational incident at FB, with values "Take-off" and "No new take-off". the flight-rate can be acquired from this variable

$\boldsymbol{xNU}$: The number of UAV units deployed
$\boldsymbol{OpDem}$: Operational demand. Two different levels have been assumed, i.e. 4/5 and 5/5 of a day air-surveillance coverage
$\boldsymbol{TOT}$: Time on Task; the realized time on task of the UAV that performs the flight
$\boldsymbol{PExp}$: The skill level of the operator (pilot) with three discrete values
$\boldsymbol{Env}$: The environmental conditions with two discrete values, “OK” and “Not OK”
$\underline{\boldsymbol{FRT\_LRU}}$: Failure Incident of a UAV due to LRU, with values “New Failure” and “No-New Failure”
$\underline{\boldsymbol{FRT\_PRU}}$: Failure Incident of a UAV due to PRU, with values “New Failure” and “No-New Failure”
$\underline{\boldsymbol{FRT\_DU}}$: Failure Incident of a UAV due to DU, with values “New Failure” and “No-New Failure”
$\boldsymbol{Rdu\_LRU}$: The duration of the UAV repair due to LRU fault
$\boldsymbol{Rdu\_PRU}$: The duration of the UAV repair due to PRU fault
$\boldsymbol{Rdu\_DU}$: The duration of the UAV repair due to DU fault
$\boldsymbol{FlHbd\_LRU}$: The number of flying hours since the last repair
$\boldsymbol{FlHbd\_PRU}$: The number of flying hours since the last repair
$\boldsymbol{FlHbd\_DU}$: The number of flying hours since the last repair
$\boldsymbol{xNM}$: The number of mechanics deployed
$\boldsymbol{MExpB}$: The skill level of the mechanic that took over the repair of the UAV
$\boldsymbol{QM}$: The percentage of mechanics that are idle
$\boldsymbol{BWkld}$: The percentage of the repair facilities that are occupied at the first-line
$\boldsymbol{xNTr}$: The number of drivers that have been deployed to do the transport from first-line to the second-line and back
$\boldsymbol{QAdm}$: The percentage of drivers that are idle
$\underline{\boldsymbol{WFRT\_LRU}}$: Workbench LRU failure Incident at the second-line, with values “New Failure” and “No New failure”
$\underline{\boldsymbol{WFRT\_PRU}}$: Workbench LRU failure Incident at the second-line, with values “New Failure” and “No New failure”
$\boldsymbol{WRdu\_LRU}$: The duration of repair
$\boldsymbol{WRdu\_PRU}$: The duration of repair
$\boldsymbol{PRUdisc}$: The mechanic’s diagnosis output about whether the PRU is repairable or not
$\boldsymbol{MExpW}$: The skill level of the mechanic that took over the repair of the component
$\boldsymbol{WWkld}$: The percentage of the second-line repair facilities that are occupied
$\underline{\boldsymbol{ORT\_PRU}}$: A PRU resupply incident, with values “New Order placed” and “No New Order placed”

$\underline{\boldsymbol{ORT\_DU}}$: A DU resupply Incident, with values “New Order placed” and “No New Order placed”
$\underline{\boldsymbol{ORT\_DP}}$: A DP resupply Incident, with values “New Order placed” and “No New Order placed”
$\boldsymbol{Odu\_PRU}$: The duration of the resupply from the moment that the ordered was placed until the item was on the depot’s shelf
$\boldsymbol{Odu\_DU}$: The duration of the resupply from the moment that the ordered was placed until the item was on the depot’s shelf
$\boldsymbol{Odu\_DP}$: The duration of the resupply from the moment that the ordered was placed until the item was on the depot’s shelf
$\boldsymbol{xSLRU}$: The nominal (order up-to) level of LRUs in the inventory
$\boldsymbol{Oh\_LRU}$: The on-hand level of LRUs
$\boldsymbol{xSPRU}$: The nominal (order up-to) level of PRUs in the inventory
$\boldsymbol{Oh\_PRU}$: The on-hand level of PRUs
$\boldsymbol{xSDU}$: The nominal (order up-to) level of DUs in the inventory
$\boldsymbol{Oh\_DU}$: The on-hand level of DUs
$\boldsymbol{xSDP}$: The nominal (order up-to) level of DPs in the inventory
$\boldsymbol{Oh\_DP}$: The on-hand level of DPs

6.4 Description of the Simulation Model’s Activity Diagram

For the reasons explained in Section 6.2, a simulation of the operations supporting a fleet of UAVs was developed.

The Support Chain is composed of three lines. In the first line of support, the UAVs that land due to a fault are diagnosed and repaired by substitution of only one of their components; therefore, it is assumed that there is only one fault per “emergency” landing and also that preventive maintenance is not included. This part of the diagram is described in more detail in Section 6.4.1.2.

The spare needed for the substitution is brought by a truck driver from the depot which is manned at the second line. In the same place, apart from the inventory there are also repair facilities for the faulty components. After the repair of a component, it is fed back into the depot’s inventory for future use. The details for this level of support are presented in Section 6.4.1.3.

Finally, there are a number of components that either due to their nature or due to their bad condition, are not repaired. Consequently, resupply orders are issued which are realised by the third line, the details of which are presented in Section 6.4.1.4.

Finally, the flight Operations themselves are simulated to include that part of the context that was found in Chapters 2 and 3 as being influential. Details for the Operations' functions are presented in Section 6.4.1.1.

The outline of the activities that were simulated is presented in Figure 6-2 with an Activity Diagram (Banks et al., 2001; Law and Kelton, 1991; Page, 1994; Shi, 1997; Visual Paradigm online, 2019). It is the Graph that was used as the plan/skeleton for the development of the Discrete Event Simulation (DES), which was coded in MATLAB.

As mentioned earlier, in the sections that follow (Sections 6.4.1 till 6.4.1.4), details are provided of the events' flows that take place at the individual lines (nodes) of the Support Chain, including the Operations, as they are presented as a whole in Figure 6-2. At the end of each section a brief revision is provided of the connections/communication that each of the lines has with the rest.

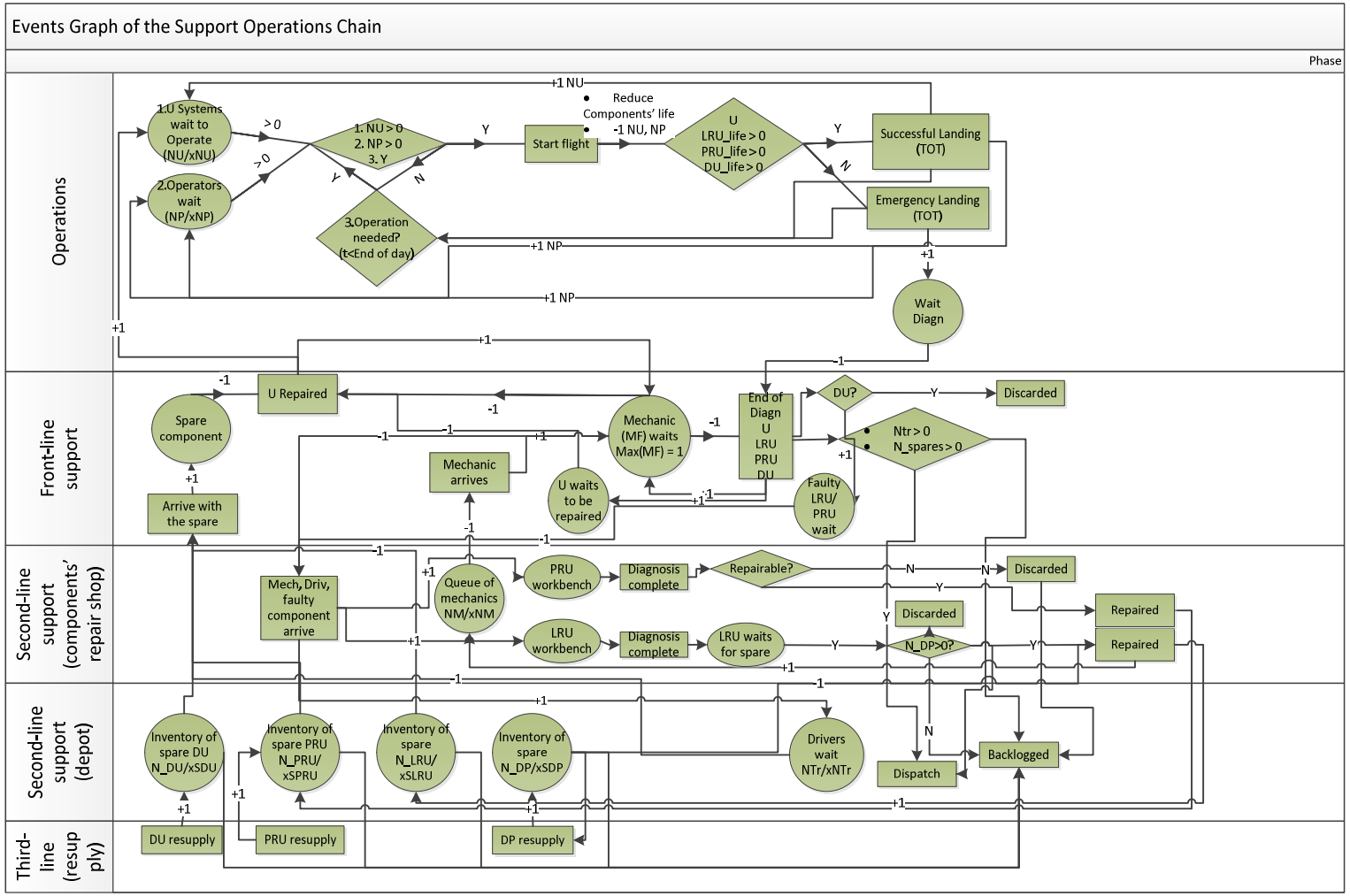


Figure 6-2: The Activity Diagram of the whole support system, including the operations

6.4.1 Activity Diagram details

6.4.1.1 Operations

The Operations vignette is presented in Figure 6-3

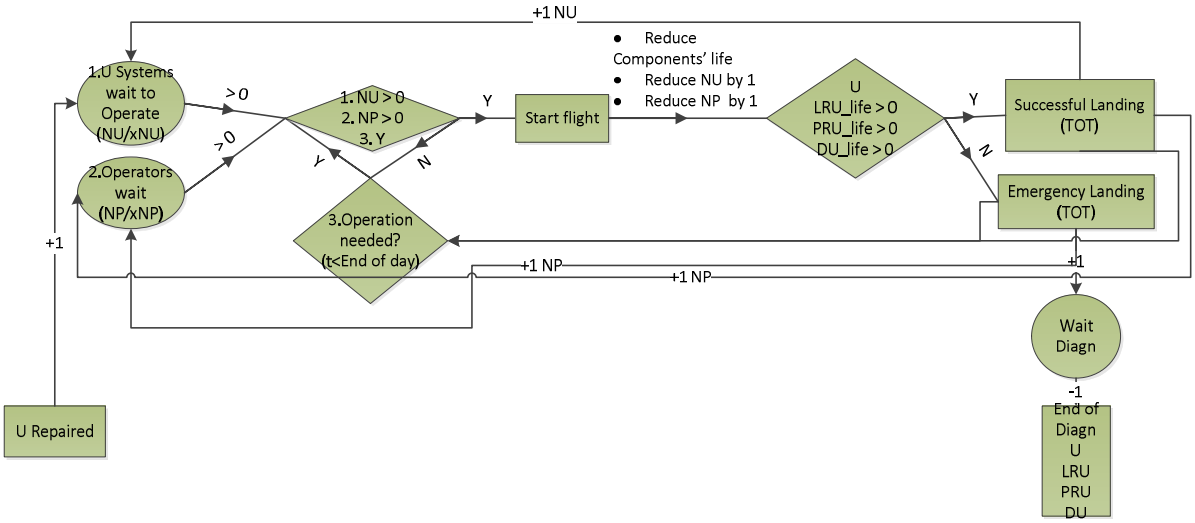


Figure 6-3: Operations

The first event that initiates the simulation model is the "Start Flight" at the "Operations" Line. It takes place if the following three conditions are all met:

- There is a requirement for a take-off (Y)
- There are systems U available (NU > 0)
- There are operators P available (NP > 0)

By the time a flight is initiated, the number of hours of the components (LRU, PRU and DU) that are on the system U are reduced. Even though it is not presented in the Activity Diagram, the amount of reduction depends on the Environmental Conditions and the Operator's skill-level/experience, and the level of reduction is different for each component in order to simulate that not all parts are affected in the same way during their operation. Furthermore, the number of systems (NU) that are available to be deployed are reduced by 1 and so are the available number of operators (NP). The initial number of systems xNU and operators xNP are decision variables and change as the scenario evolves to each of the different Phases (see Table 7-2, Table 7-10).

There are two possible Events that can follow. The first one is that the mission is fully performed by the system – a “Successful Landing” event - and the system U gets back to fill the pool of the available systems. Consequently, the number of systems available are increased (+1 to the “U Systems wait to Operate” queue) and so are the number of operators (+1 to the “Operators wait” queue). One detail that is worth mentioning is that if the operational demand is low – i.e. not for the full duration of the 24hrs - and the flight’s time is close to the end of the day and thus no systems need to operate, the model checks this condition and lands the system. This check has not been included in the Graph to avoid further clustering. Finally, a check is performed if another “Start Flight” event needs to take place.

The second possible event has to do with a fault. The duration asked to be flown, modified by the effect of the Environmental Conditions and the skill-level of the operator is checked against the residual hours of each component of the system and if in one of them the modified duration exceeds the hours left, then there is an “Emergency Landing” event. When such an event happens, the operator returns (“Operators wait” queue) and the number of available operators is increased by 1. Furthermore, a check is made if the mission needs to be covered by another system, so as to have another “Start Flight” event. The system that has malfunctioned then waits for a mechanic to diagnose the problem (“Wait Diagn” queue).

The Operations’ vignette communicates directly with the First-line support in two ways. Firstly, it sends out. The “Wait Diagn” queue is one of the inputs for the “End Diagn U” event. Secondly, the Operations receive from the “U Repaired” event of the First-line, any repaired systems into the “U Systems wait to Operate” queue.

6.4.1.2 First-Line Support

The First-line support vignette is presented in Figure 6-4.

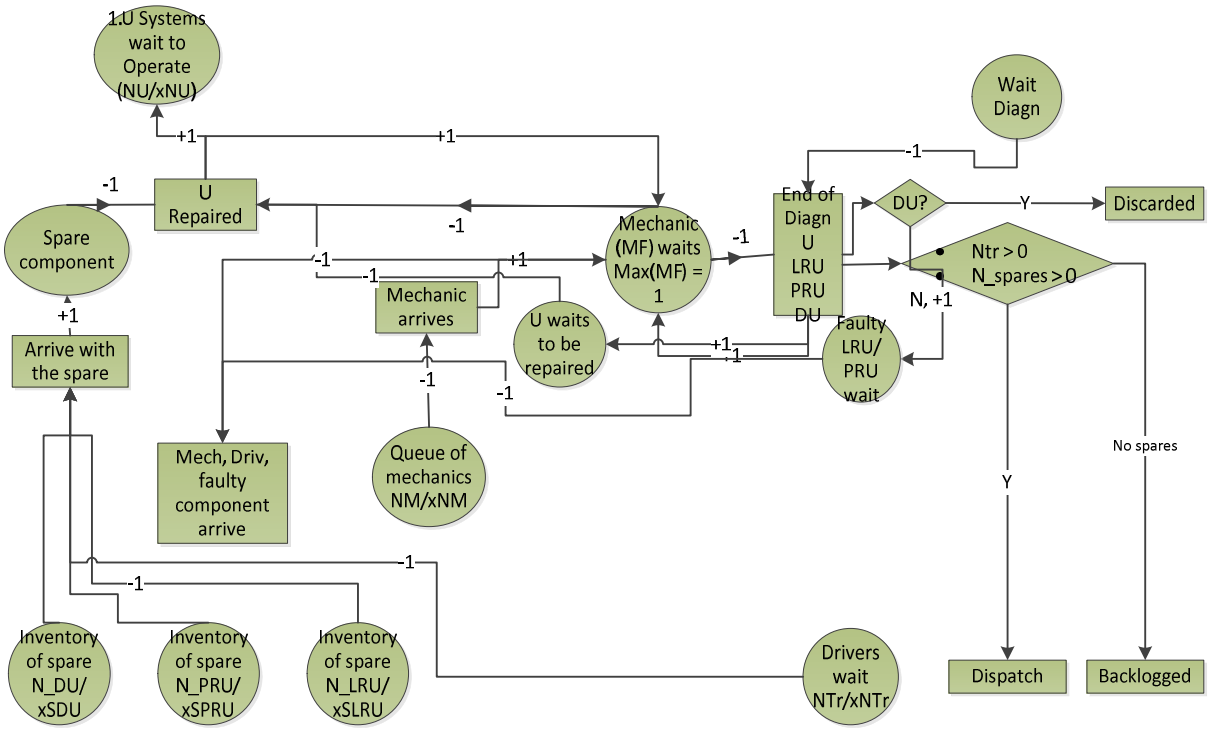


Figure 6-4: First-line support

If a mechanic is available from the “Mechanic (MF) waits” queue, then the malfunctioned system enters the area of the “Front-line of support” and there is the “End of Diagnosis” event. The duration of this event depends on the skill-level/experience of the mechanic. If a mechanic is not available in the queue, the system waits for one to become available from the Second-line support (components’ repair shop) where all the mechanics stay (“Queue of mechanics”). The initial number of which xNM is a decision variable. This is signified in the Graph via the “Mechanic arrives” event and then the mechanic is directed to the “Mechanic (MF) waits” queue. After the diagnosis is complete – and this is signified by the “End of Diagn U” event - an order is sent to the Second-line support (depot) to provide a spare component so as to replace the faulty one.

If the faulty component in the system is a DU then it is discarded (a “Discarded” event), but a resupply order is not initiated yet, until a new one is dispatched from the inventory. If the fault was either due to an LRU, or a PRU, then the faulty component is removed from the system and the component waits in the “Faulty LRU/PRU wait” queue in order to be taken back in the repair shop, while the system waits in the “U waits to be repaired” queue. Furthermore, the mechanic

waits in the “Mechanic (MF) waits” queue, for the new spare to arrive from the Second-line. The order is received by one of the drivers through the “Dispatch” event. The drivers wait in the “Drivers wait” queue at the “Second-line support (depot)” area. If no spare is available at the moment, then there is a “Backlogged” event.

At the First-line support area, the “Arrive with spare” event takes place through which a driver brings the required spare component. The duration of the transportation depends on the Environmental Conditions.

The component then sits in the “Spare component” queue. There also is a system in the “U waits to be repaired” queue, so if a mechanic is available in the “Mechanic (MF) waits” queue, a repair activity leads to the “U Repaired” event. The duration of the repair activity depends on the skill-level/experience of the mechanic. The “U Repair” event then leads to having one more system available for the “U Systems wait to Operate” queue at the “Operations” area, with one of the system’s components in new condition, while the others have the previously accumulated hours.

After the “U Repaired” event, the mechanic and the driver are made available. The mechanic notionally goes back to the “Mechanic (MF) waits” queue, but then immediately, along with the driver and the faulty component that sits in the “Faulty LRU/PRU wait” queue, drive to the “Second-line support (components’ repair shop)” and the “Mech, Driv, faulty component arrive” event that takes place. The duration of the travel depends on the Environmental Conditions.

The First-line support vignette communicates directly with the Operations, the Second-line support (components’ repair shop) but mostly with the Second-line support (depot) area in order to get the required spares. Regarding the Operations, it receives the faulty system from the “Wait Diagn” queue, and sends back a repaired system to enter the “Systems wait to Operate” queue. The communication with the Second-line support (components’ repair shop) is again not very dense.

The First-line support receives a mechanic from the “Queue of mechanics” which is at the Second-line support (components’ repair shop). The assumption is that

if there is one available in the queue then the dispatch to the First-line, which is right next to the Operations, is done automatically, without any request. This mechanic then stands-by to be ready to deal with any problem. The First-line again, sends out to the Second-line support (components' repair shop) the faulty component accompanied by the mechanic who removed it and the driver who transports them. The event that takes place is the "Mach, Driv, faulty component arrive" event.

As mentioned earlier, the communication with the Second-line support (depot) is denser. It is there that a "Dispatch" order event takes place, or if no parts are in the inventory, a "Backlogged" event is recorded. From that area the four queues of the three types of components, that can probably be used for the repair of the system by substitution at the Front-line (LRU, PRU or DU), and of the drivers, are fed into the "Arrive with the spare" event. These queues are the "Inventory of pare N_DU", "Inventory of spare N_PRU", "Inventory of spare N_LRU" and "Drivers wait" respectively.

6.4.1.3 Second-Line Support: Repair Shop

The Second-line support (components' repair shop) vignette is presented in Figure 6-5.

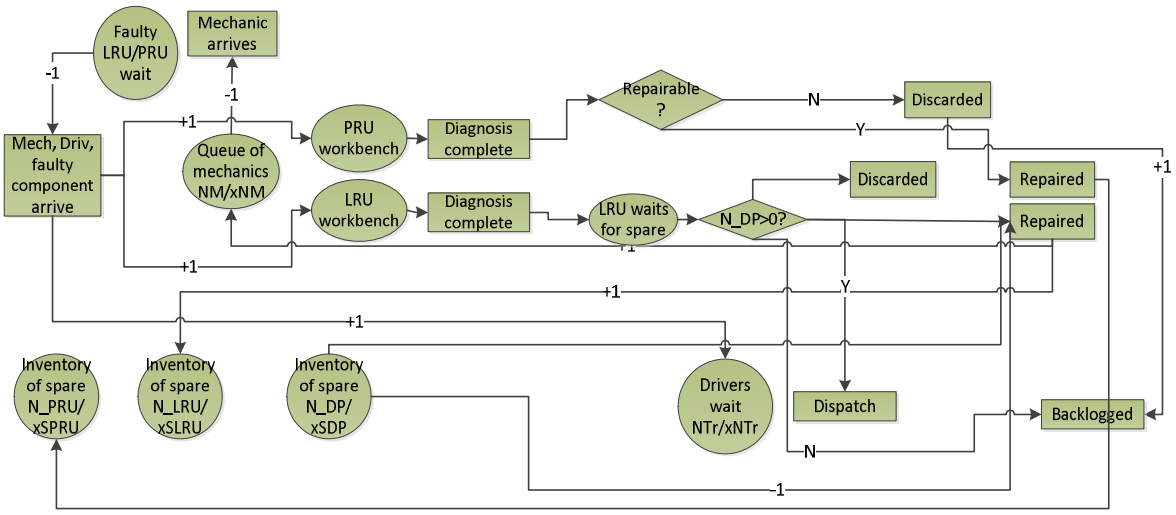


Figure 6-5: Second-line support (components' repair shop)

At the Second-line support (components' repair shop) it is the "Mach, Driv, faulty component arrive" event that initiates the processes that take place there. Before

discussing how this has been modelled, it should be mentioned that after the arrival, the driver is released to go back to wait in the “Drivers wait” queue at the Second-line support (depot) and also that the repair process of the component is performed by the same mechanic that brought it to the shop. The reason that for the assumption that the same mechanic performs the repair of the faulty component, is that there is also the assumption that no batching in the repairs (or the resupply orders) is allowed.

If the component is a PRU, a diagnosis firstly takes place and thus there is a “Diagnosis complete” event. The diagnosis reveals if the PRU component is repairable. If it is, then a “Repaired” event takes place. The “Repaired” event then fills the PRU inventory (“Inventory of spare PRU”) at the Second-line support (depot) with one more part and also the mechanic is released and goes back to the “Queue of mechanics”. On the other hand, if the faulty PRU is beyond repair, then there is a “Discarded” event. This event triggers a “Backlogged” event at the “Third-line (resupply)”. Finally, the “Discarded” event releases the Mechanic who then is added to the “Queue of mechanics”.

If the component is an LRU, a diagnosis firstly takes place again, and thus there is another “Diagnosis complete” event. This event places the under repair LRU in the “LRU waits for spare” queue. The diagnosis (always) reveals that the LRU needs a DP part to be repaired, which is located in the inventory that is called “Inventory of spare N_DP” queue which is at the “Second-level support (depot)”. If the shelves are not empty, then one is dispatched. Finally, as above, the “Repaired” event adds one more LRU spare in the “Inventory of spare N_LRU” queue at the “Second-level support (depot)”, and the same event also releases the Mechanic who then is added to the “Queue of mechanics”. The replaced DP is discarded (“Discarded” event).

As before, the duration of the diagnosis and of the repair activities depend on the skill-level/experience of the mechanic. Furthermore, the duration of the transportation of the faulty component from the First-line to the Second-line support (components’ repair shop) depends on the Environmental Conditions.

The Second-line support (components' repair shop) vignette communicates directly with the First-line support and the Second-line support (depot) area in order to get the required spares but also to fill in the inventory with the repaired components.

The Second-line support (components' repair shop) receives the "triplet" of the faulty component, the mechanic and the driver from the "Faulty LRU/PRU wait" queue and provided the "Mech, Driv, faulty component arrive" event. Furthermore, the Second-line support (components' repair shop) feeds back to the First-line a mechanic from the "Queue of mechanics" when one becomes available, and thus the "Mechanic arrives" event takes place.

The Second-line support (components' repair shop) feeds the PRU and LRU inventory of the Second-line support (depot) with spare parts through a "Dispatch" event. As it was shown, the inventory is represented in the Graph via the "Inventory of spare N_PRU" and "Inventory of spare N_LRU" queues respectively. The "Drivers wait" queue of the drivers at the Second-line support (depot) is also fed with one more after the "Mech, Driv, faulty component arrive" event takes place. The "Inventory of spare N_DP" from the Second-line support (depot) provides one part so the "Repaired" event for the faulty LRU can take place at the Second-line support (components' repair shop). If a part is not available on the shelf then a "Backlogged" event takes place at the Second-line support (depot) area. A similar backlogging action happens if the diagnosis of a faulty PRU shows that it is beyond repair and a "Discarded" event takes place at the Second-line support (components' repair shop).

6.4.1.4 Second-Line Support: Depot and Third-Line (Resupply)

The Second-line support (depot) along with the Third-line (resupply) vignette is presented in Figure 6-6.

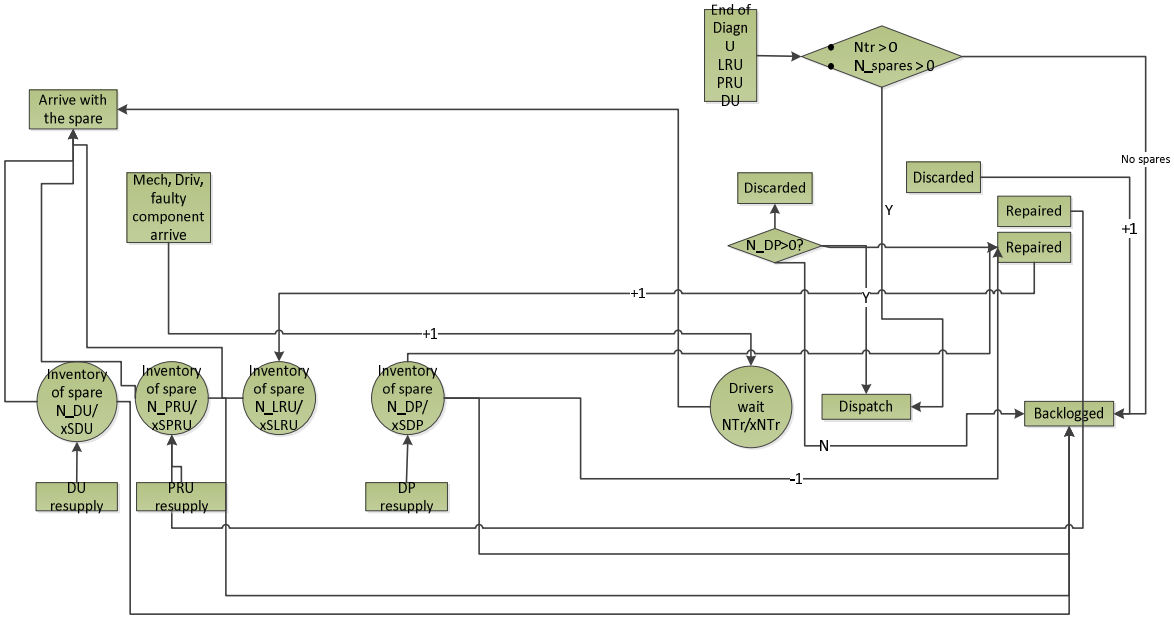


Figure 6-6: Second-line support (depot) and Third-line (resupply)

The Second-line support (depot) is the area where the Depot lies with the inventory with spares of the four different component parts (LRU, PRU, DU and DP) and which they are represented in the Graph as queues. Furthermore, this is also where the dispatching and the backlogging are managed which is represented through the “Dispatch” and the “Backlogged” events respectively.

The “Inventory of spare DU” queue, or the “Inventory of spare PRU”, or the “Inventory of spare LRU” is reduced by 1 after the “End of Diagn U” event from the First-line support shows that a DU, or a PRU or an LRU is needed respectively, and a check that there are available drivers in the “Drivers wait” queue shows a driver’s availability. This action takes place through the “Dispatch” event.

Given that there is the assumption that the components’ unit costs are high, then a resupply inventory policy of (S, S-1) is justified. Consequently, by the time a unit is removed from the shelves, another is backlogged via the “Backlogged” event and also ordered from the “Third-line (resupply)”. After some time, which is affected by the Environmental Conditions, a respective resupply event fills the inventory.

Regarding the DU, the “Backlogged” event is controlled from the First-line support area and the “Inventory of spare N_DU” is resupplied after a “DU_resupply” event takes place from the Third-line. However, this is not exactly the same with the other three types of spares.

As was seen earlier, the backlogging of the PRU and the DP is not initiated from the First-line, but from the Second-line support (components’ repair shop) area. This is because the PRU is backlogged only if it is judged beyond repair and thus is discarded (“Discard” event in the Second-line support (components’ repair shop)). From the same area though, those PRUs that are repaired are fed then back directly into the “Inventory of spare N_PRU”. Of course, this queue is also resupplied after a “PRU_resupply” event from the Third-line area.

Regarding the DPs, they are used at the Second-line support (components’ repair shop) area too, in order to repair a faulty LRU. A dispatch order is initiated there and if a new one is available in the inventory then it is dispatched and a resupply order is issued (“Backlogged” event). If a DP is not available, then a resupply order is issued as well. The “Inventory of spare N_DP” queue is replenished after a “DP_resupply” event takes place from the Third-line.

The LRUs are assumed to be always repairable and thus are never discarded. This means that there is no backlogging for them. The only way that the “Inventory of spares N_LRU” is filled is after a repair with a “Repaired” event takes place at the Second-line support (components’ repair shop) area.

The Second-line support (depot) is the area where the queue with the drivers is (“Driver wait” queue). It sends out a driver in order to participate in the realisation of the “Arrive with the spare” event at the First-line and then it is fed back with a driver when the “Mech, Driv, faulty component arrive” event takes place. The drivers’ nominal number is $xNTr$ and it is a decision variable and so are the target values for the spares $xSLRU, xSPRU, xSDU$ and $xSDP$ in their respective inventories.

In order to simulate that the resupply from the manufacturer is a long-leg activity within the Supply Chain, the duration of the resupply at the Third-line takes a comparatively long time and also depends on the Environmental Conditions.

6.5 Conclusions

In this chapter, the third conceptual model presented in Chapter 3 was used as a framework to develop a simulation of the Support Chain (SC). Furthermore, details were presented of the Activity Diagram that was used as the skeleton for the computer simulation model.

That simulation was then used to generate the required data for model development and evaluation in Chapter 7.

7 RESULTS AND DISCUSSION

7.1 Introduction

Two simulation studies were performed of the same system under different scenarios. The first one was intentionally simpler than the second, so that it was possible to get results from a simple model and then consider how they would be transferred and expanded into a comparatively more elaborate model. The difference was that in the simple scenario there were only two components of interest: the LRU that can make the UAV stop operating and the DP that then was required to be replaced in the LRU, while in the second scenario all the components that have been described in Section 6.4 could malfunction. The Activity Diagram (AD) of the simulation is presented in Figure 6-2.

A description of each scenario now follows, along with the demand forecast models that were used and their evaluation.

7.2 Simulation Support Scenario – Case 1

The simulation (see Figure 7-1 and the first part of Section 6.4) concerned the support provided to a small fleet of generic Unmanned Aerial Vehicles (UAVs) that was used for ISTAR operations at a single Forward Base (FB). Their Logistics Support Organisation (LSO) was composed of a Forward support level (FORWARD) at which failed components that could make the UAVs non-operational (it was assumed only LRUs for Case 1) were replaced with new ones from inventory, and a Central repair level (CENTRAL) at some distance from the FB where the inventory of spares were kept and repairs were performed on the failed LRU components.

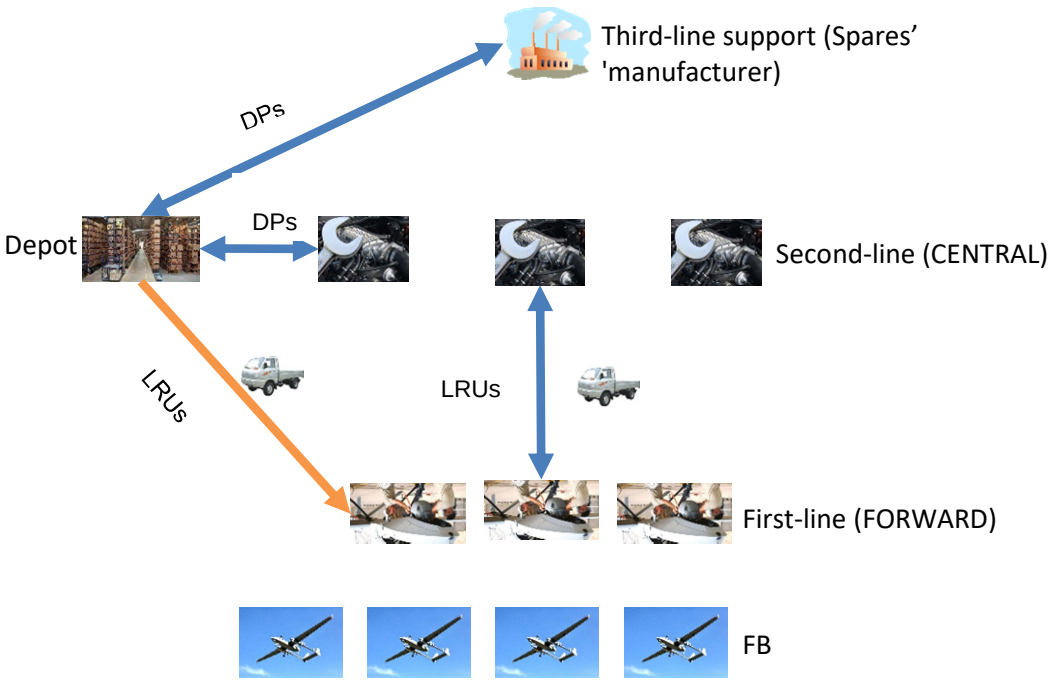


Figure 7-1: The simulated Supply Chain and Operations

In this Case 1 scenario, the Equipment Breakdown Structure (EBS) of a UAV unit was composed of only a single LRU that could be repaired at the CENTRAL depot by the replacement of a single Disposable Part (DP) kept in the same store as the LRUs. Furthermore, it was assumed that systems' innate failure rates did not change with age and only corrective maintenance was applied, which means that preventive maintenance policies were not considered.

In the assumed scenario, each UAV had a nominal Time on Task (TOT) of four hours, after which it had to land for a quick refuelling. If another UAV was available then it took off; if not, the same UAV took off again.

The operational demand was for a single UAV unit to cover an area assigned for ISTAR operations during a given proportion of the day, each day. For example, if the operational demand was to cover 4/5 of the day, since either there was no need to fly during night hours, or a different group took over that period, then the operational demand ($OpDem$) was "4/5". Because of the importance of the ISTAR functions, any available, unassigned mechanic was assumed to be waiting in FORWARD to help in case of a breakdown (B).

Assuming the occurrence of a breakdown, another UAV would take off if one was available. Subsequently, the grounded UAV was taken over by a mechanic who then started the diagnosis procedure. The duration of this procedure depended on the skill level of the mechanic, and it was assumed that the fault was always a single one and was always found correctly. After the diagnosis was over, an order for a spare was given to the CENTRAL depot. The spare took some time to be located and acquired by a driver and was then brought to FORWARD. The mechanic would replace the faulty LRU with the spare, making the UAV available again. The faulty LRU was then transported back to CENTRAL by the mechanic and the driver in order to be repaired.

There were three available workbenches (Ws) at CENTRAL which were used for diagnosis and repair of the faulty items. Due to the importance of the part, there was no batching in the repair activities. Therefore, the same mechanic was assumed to undertake the diagnosis and repair on one of the available workbenches and bring the LRU in a usable condition back to the LRU inventory, provided there was a DP in stock. Due to the assumed high cost of a DP, the depot was using an (S, S-1) inventory policy and thus initiated a resupply order whenever there was a single DP unit removed from the DP inventory.

In case any of the resources were not available, the related activity waited. So, for example at a specific moment if all spare LRUs were under repair at the CENTRAL, then any broken down UAV would wait FORWARD until a spare one would become available, or when all drivers were occupied, no transports of parts and mechanics could take place until a driver was released.

A snapshot example of the data collected is presented in the following table:

Table 7-1: Snapshot example of the data collected from the simulation

Incident	Time (hours from simulation start)	Object ID
StartFlight	196	UAV2
Landed	200	UAV2
StartFlight	200	UAV1
Landed	204	UAV1
StartFlight	204	UAV2
UAV broken down	204.6577	UAV2
Mechanic starts the diagnosis	204.6577	UAV2
StartFlight	204.6577	UAV1
Landed	208.6577	UAV1
Mechanic finishes the diagnosis – spare LRU required	206.787	UAV2
Mechanic starts LRU replacement on UAV	210.4336	UAV2

7.2.1 Scenario for Dataset Generation

The chosen scenario involved a single iteration of the following eight consecutive phases (Table 7-2):

Table 7-2: Scenario Phases

Phase	Duration (Months)	$xSLRU$	$xSDP$	xNU	xNM	$xNTr$	$OpDem$ (ratio of a day)	Env = OK (prob)
1	3	3	3	2	2	1	4/5	0.6
2	3	3	3	3	3	2	4/5	0.6
3	4	4	5	4	3	3	4/5	0.6
4	3	4	6	3	2	3	4/5	0.6
5	3	3	3	2	2	1	5/5	0.6
6	3	3	3	3	3	2	5/5	0.6
7	4	4	5	4	3	3	5/5	0.6
8	3	4	6	3	2	3	5/5	0.6

The assumed story behind the phases shown above was that during the 1st phase when operations started, there were two UAVs (xNU = 2) deployed with a mission

to provide ISTAR functions for the Operational Demand ($OpDem$) of 4/5 of a day. For the manning of the LSO in the 1st phase, there were two mechanics deployed (xNM = 2) and one driver ($xNTr$ = 1), while the initial spares stock levels were three LRUs and three DPs ($xSLRU$ = 3, $xSDP$ = 3). The UAVs were flown by an equal number of operators with an initially sampled level of proficiency.

As the operations were seen to continue with an anticipated future increase in the need for air-surveillance, in Phase 2 an additional UAV was deployed along with an additional driver to help with the transports of the spares and the mechanics. This situation lasted for three months and was followed by Phase 3, a four months phase when a further 4th UAV was deployed. The target levels of spares of LRUs and DPs were also increased at the beginning of Phase 3.

In Phase 4, one UAV was withdrawn along with a mechanic. In Phase 5, the $OpDem$ had to be increased to full 24hrs surveillance (5/5), although at the same time, one UAV was assumed to have failed beyond repair. In addition, it was assumed that two drivers were transferred out of the LSO, while the target level of spares was reduced. Further changes of this nature affecting the LSO's configuration were assumed for Phases 6 to 8, as shown in Table 7-2. Finally, the environmental conditions throughout these phases were assumed to be good (Level 1 / "OK") with a probability of 60%.

Records of incidents (e.g. take-offs and landings, of break-downs, of repair and re-order), of levels of resources (spares, mechanics, drivers) and of durations (hours flown, repair and resupply times) were kept from the single run of the eight consecutive phases, just like the records that would be kept in the relative logs of real operations. Furthermore, variables that can affect the incidents and the duration of diagnosis, repair and transport were also recorded. Such variables were the environmental conditions, the operators' skill levels/ experience, the mechanics' skill level / experience and their workload level.

7.2.2 Simulation of Test Data to Allow Forecast Comparison

The end of Phase 8 provided the initial conditions for a follow-on ninth phase of six months' duration that was used to evaluate the performance of the demand prediction models. The research of interest has been in how well can demand

predictions be provided when the failure-context factors are about to change during the final-phase. Consequently, Phase 9 - the final phase (FPP) - could take different courses in order to represent a range of changes likely to be experienced in practice.

The following table summarises the full factorial experimental design of the contexts which were simulated for 100 repetitions each in order to provide the data needed to evaluate the forecast models. The design produced 144 different contexts:

Table 7-3: The combinations of Phase 9 configurations that constituted the test dataset

	$(xSLRU, xSDP)$	xNU	xNM	$OpDem$ **(ratio of a day)**	Env **= OK (prob)**
Phase 9	(3, 3) , (4, 5) , (4, 6) , (8, 8)	2 , 3 ,4	2 , 3	4/5 , 5/5	0.3 , 0.5 , 0.7

7.2.3 Forecasting Approaches Employed

A main objective of the present research (Section 1.4) has been to use the factors that could be elicited from the logbook records that were kept during the building-up and infinite-time horizon phases 1 to 8 in order to model the context of the demand for spares. The intention was to use any information that could be available at the beginning of the final-phase 9 in order to forecast the demand during that phase. The assumption was that what could be known in advance is what is presented in Table 7-3.

The main value of interest was the probability of experiencing a failure incident in any given hour (binomial variable FRT_LRU in Table 7-2, which for simplicity in this scenario was called FRT). It was possible to derive the required mean number of failures for the duration of the forecasting period by multiplying the acquired FRT by the respective 4,320 hours included in the 6 months of the final phase 9. It was assumed that the operated systems do not degrade. The

engineering context was not examined since the assumption was that the same UAVs operated throughout the study.

The other core objective of this research has been to evaluate the applicability of BNs under different ways of developing their structure (Section 1.4) for the forecasting of the demand during the final-phase 9. A BN can be developed in different ways, using different combinations of human expertise and data (Korb and Nicholson, 2004). It can be developed entirely from a dataset and this was what the present research firstly examined. This entails both the structure of the network, i.e. the DAG, and the associated NPTs being derived from the dataset. While obtaining NPTs from a dataset is relatively straightforward, deriving the structure is much more involved. The reason is primarily due to the huge number of DAGs which can be built from even a relatively small number of variables. Nevertheless, for the reasons that were discussed in Section 4.3, it was found that score based structure learning provides a lot of benefits and this method was applied here.

Instead of deriving a BN's structure from data, another common approach is to elicit the structure from a subject matter expert (SME). Such a DAG is usually easier to understand and therefore more easily explained and comprehended by decision makers.

However, due to the complexity of the situation, there might be connections that are missing from an expert-elicited structure. In such a case a hybrid approach can also be adopted. Here, the subject matter expert can provide an initial DAG which is then built upon by an automated machine learning algorithm. The aim of such an approach is to ensure that key relationships are preserved and that more subtle effects are not missed.

Furthermore, the expert-elicited structure can also be used as a starting solution for a score-based algorithm that then starts searching the solutions' space by adding, removing or reversing arcs among the variables.

All the BN models were built in R, using the bnlearn package (Nagarajan, Scutari and Lebre, 2013).

In order to provide a comparison with the BN predictions coming from the four different BNs, forecasts using two other methods were also provided. The first was a logistic regression, which also tried to account for the relationships between the contextual factors and the observed number of failures. The appropriateness of this type of regression model stems from the underlying random process which involves the generation of failed equipment. The output, as for the BNs, was the probability of experiencing a failure incident in any specific hour.

The second type of additional forecast was human judgement. Along with the starting configuration for the ninth/final operational phase, the judges were also supplied with the Single Exponential Smoothing (SES) forecast available at the end of the eighth operational phase. This can be described as an expert adjusted forecast, with adjustment being made away from the fixed SES forecast.

The evaluation was performed using both accuracy metrics and accuracy implication metrics. Regarding the accuracy metrics, the Mean Absolute Scaled Error ($MASE$) was applied but adjusted by taking the geometric mean over the different 144 datasets for the reasons that were mentioned in the respective Section 5.2.3.3. As far as the accuracy implication metrics are concerned the following five measures were used:

1. The average volume of inventory during the period
2. The volume of spares left at the end of the period
3. The probability of no stock-outs during the period
4. The system's average Operational Availability for the period of interest
5. The systems' Operational Availability at the very end of the period

7.2.3.1 BN Learnt from Data

Using the BN learning package in R called "bnlearn" the sampled dataset of records from Phases 1 to 8 was fed into a score-based unsupervised learning algorithm. The scoring method employed the Modified Bayesian Dirichlet equivalent uniform (MBDeu) score (Cooper and Yoo, 1999; Heckerman et al., 1995). As discussed in Section 4.3.2.2, an optimisation algorithm such as tabu,

was applied to search for the structure that could score highest in the MBDeu. The search algorithm starts by having a random network, and then proceeds by adding and removing connections among the variables/nodes. The inclusion or removal of connections aims to increase the value of the MBDeu score. The searching process was applied to 300 bootstraps and thus developed 300 networks that were averaged to form the final network.

The above procedure produced the network displayed in Figure 7-2.

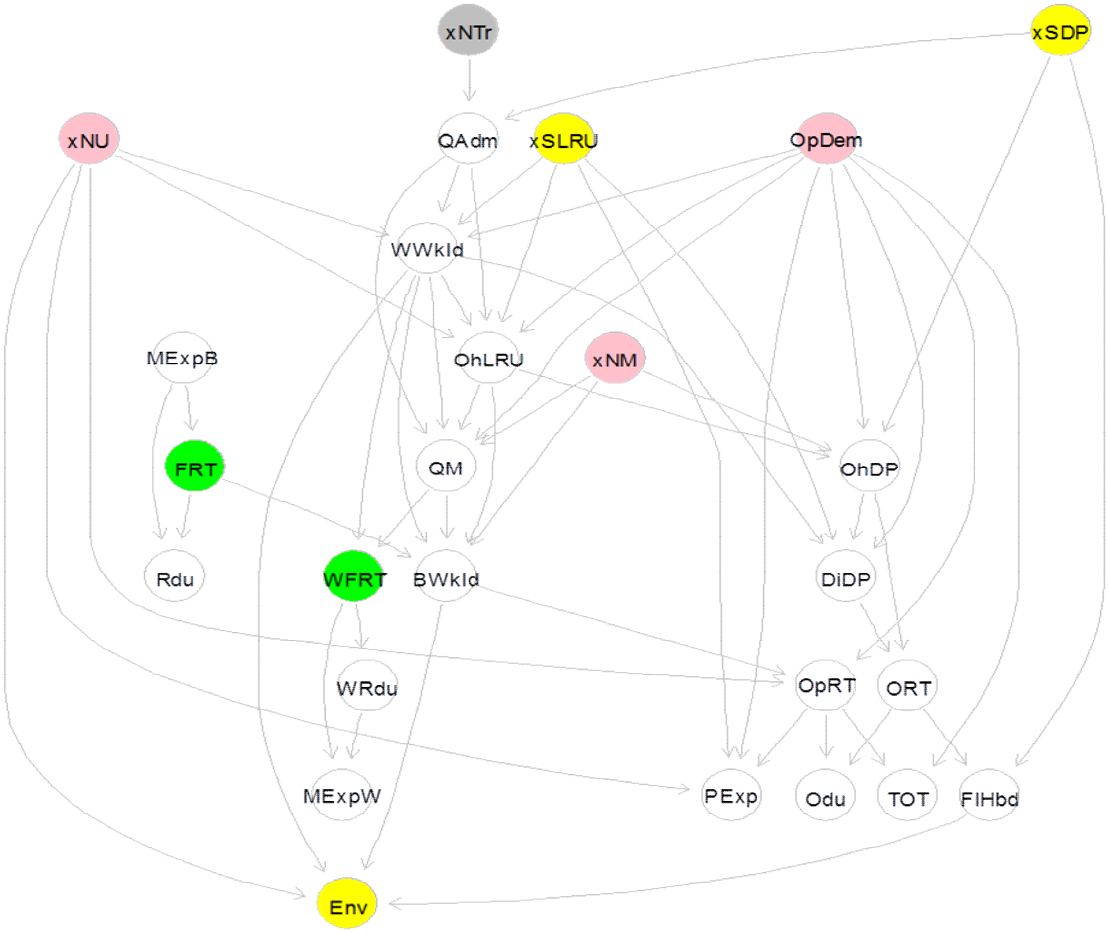


Figure 7-2: DAG of the BN model that was learnt from the simulation training dataset

Note that the resulting model is not a causal BN since the causality assumptions are not met (see e.g. Pearl (1988)). However, it does provide an interpretation of the relationships / associations among the variables. For example, the arc which connects xNU directly to $OpRT$ and the arc that connects the latter to the TOT

indicate that the number of units operated (xNU) has a direct effect on the Operational Rate ($OpRT$), and that on the TOT, which means that the number of UAVs deployed is associated with how often take-offs are missed or performed, and that has an effect on the duration of any single take-off (TOT).

Furthermore, most of the arcs are directed towards the variables $OhLRU$ (the on-hand LRU), $WWkld$ (how busy the repair workshops are at the CENTRAL level) and $BWkld$ (how busy the workshops are at the FORWARD level). This indicates that these facilities are key to the whole system. Finally, many arcs start from the $OpDem$, which indicates that this is another key factor to the context.

7.2.3.2 Expert-Elicited BN

A BN of the problem situation was developed by eliciting a DAG from a domain expert, who was one of the participants of the interviews discussed in Section 3.2. This DAG displays the relationships believed by the expert to exist in the system, using the idioms that were described in Section 4.3.2.5.

However, the fact that the primary intent was to be able to use the variables that would be available from logbook records, created restrictions of how the variables' relationships can be expressed. A failure, an end of repair, the condition of the environment, etc. are incidents that have been recorded in the logbooks at a specific time-instance. Consequently, when considering the relationship of the Environment to the Failure, in a more common case, one would apply the cause/consequence idiom and thus, connect an arrow from the Env node to the FRT node to indicate that a harsh environmental condition would cause/make a failure incident more probable. However, when the datum was recorded in an instance-incident form, the values that were acquired were not necessarily a result of a causal mechanism: a bad weather in the same record as a failure incident, has not contributed to the cause of that failure even though it is counted as such in the calculations of the NPTs. In such cases it could be said that the relationship between the two variables could be the one suggested by the measurement/indicators idiom: the presence of the bad weather indicates that a failure is more probable to appear at the same time. On the other hand, if the bad weather has been one of the causes for the failure they would have existed

at the same time as the failure appeared and thus, the causal relationship exists. Similar problems can be expressed for the relationship of the operator's experience ($PExp$) and the FRT.

In order to resolve this ambiguity, it was decided to try and maintain the intuitive causality whenever that was possible. The DAG elicited from the domain expert is presented in Figure 7-3.

As can be observed, Figure 7-3 is different to Figure 7-2. They are different because they have been built using different methods and having different assumptions. On the one hand, Figure 7-2 maximises the MBDeu score metric by the use of certain assumptions. In particular, the assumption of *likelihood equivalence* (Assumption 6, Section 4.3.2.2), as discussed in Section 4.3.2.4 maintains only the associative relationship among the connected nodes. On the other hand, Figure 7-3 has been built using the techniques described in Section 4.3.2.5 (Sections 4.3.2.5.1, 4.3.2.5.2 and 4.3.2.5.3), These techniques try to use the domain knowledge and in this way to preserve the understanding of the SME on the conditional probability relationships among the nodes.

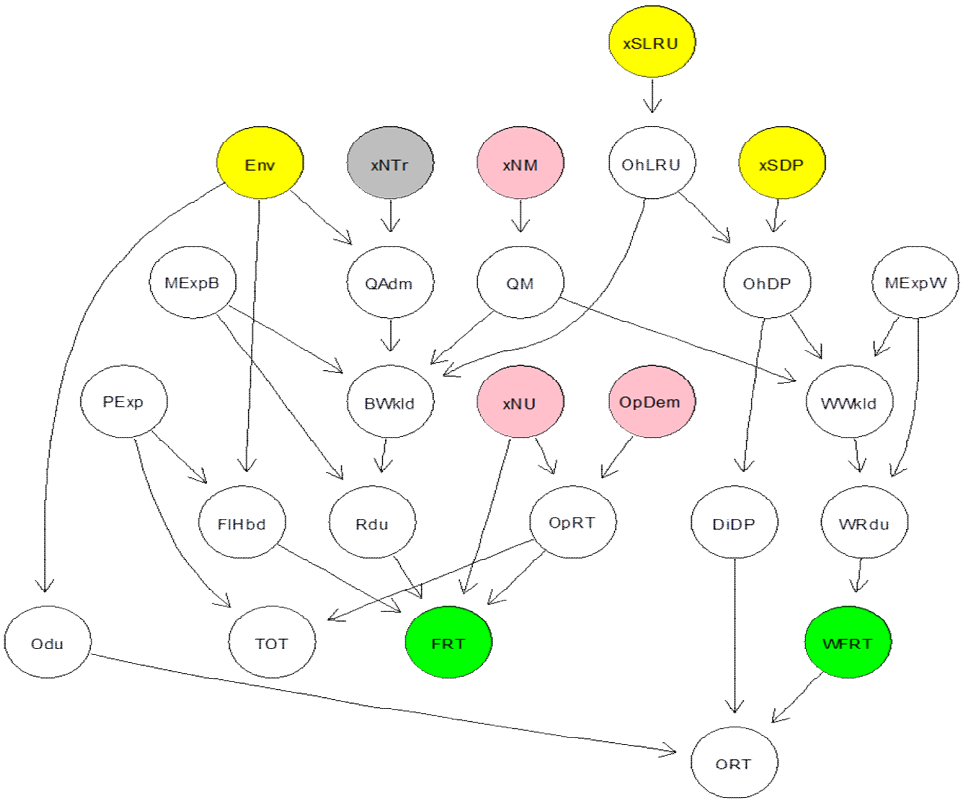


Figure 7-3: DAG of a BN model elicited from a domain expert

7.2.3.3 Hybrid BN that Maintains the Elicited Structure (BN hybrid 1)

A hybrid BN was developed in order to combine the understandable nature of the expert-elicited BN combined with the ability to learn less obvious relationships provided by the automated BN. The development of this hybrid BN began with a simplified version of the expert-elicited BN and used this as a starting point for the machine learning algorithm which was employed to develop the learnt BN. The resulting structure constrains the final DAG to incorporate the expert-elicited components but allows additional relationships to be included. A simplified version was chosen of the elicited and not the actual, primarily so as not to restrict the tabu algorithm's search area: starting from the same, very restricting point reduces the search space.

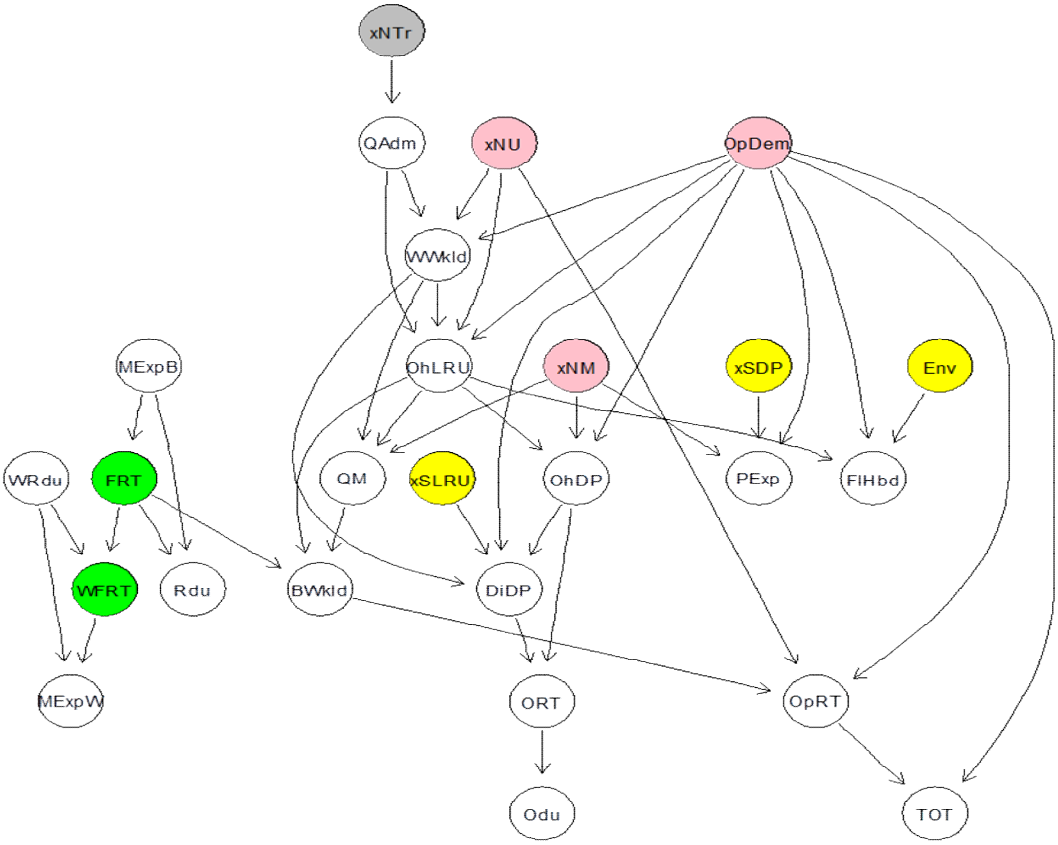


Figure 7-4: DAG of a hybrid BN, combining expert elicitation and machine learning (BN hybrid 1)

7.2.3.4 Hybrid BN that Starts from the Elicited Structure (BN hybrid 2)

A further hybrid BN was developed in order to take advantage of the fact that the elicited graph could work merely as a more suitable starting point for the tabu search algorithm instead of from the random starts that the conventional unsupervised method does. The algorithm would then proceed without any user-imposed restrictions to the structure, as was done in the previous hybrid case.

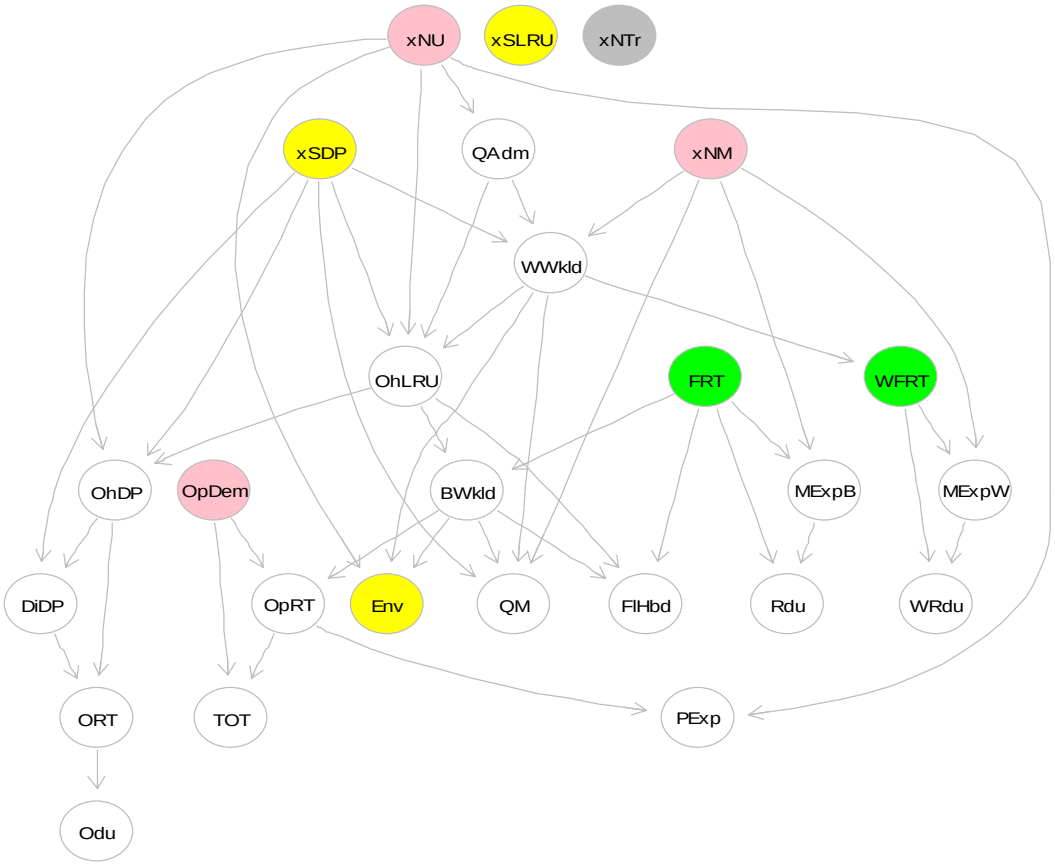


Figure 7-5: DAG of a hybrid BN, starting from the expert elicited structure and then applying machine learning (BN hybrid 2)

7.2.3.5 Logistic Regression Model

The logistic regression model derived from the first eight phases of the simulation training dataset was the following:

$$logit(FRT) = b_0 + b_1 OpDem + b_2 Env$$

, where FRT corresponds to the occurrence of an equipment failure, $OpDem$ represents the level of operational demand and Env represents the severity of environmental conditions. The model was developed using the backward variable entry method and by verifying the predictability of the model through leave-one-out cross validation.

The resulting coefficients of b_0, b_1 and b_2 were -4.5273, 0.4418 and 0.1836, respectively, with standard errors of 0.1276, 0.1212 and 0.1274. The reference settings of the variables were '4/5 of a day' for the $OpDem$ and 'OK' for the Env.

In order to forecast demand for Phase 9, where the state of the Env variable is not yet known but there is a probability distribution for it, the forecast used the probability values as weights for a weighted average of the two outputs obtained using the two possible values for the Environment.

7.2.3.6 Expert-Elicited Forecast

In order to construct this forecast, four domain experts were consulted. Each was talked through the scenario implemented in the simulation and provided with the same information. This consisted of the configurations of the eight initial phases of operation and the resulting number of failures observed in each. Every expert was then asked to provide a forecast of the number of failures expected for a final ninth phase of operations given the LSO configuration and the Single Exponential Smoothing (SES) estimate. The fixed SES forecast was obtained using the "tsintermittent" R-package which was trained with monthly demand data and used a smoothing factor of 0.2.

Due to experts' time and attention-span limitations, 18 different possible configurations for Phase 9 (Table 7-4) were sampled and each of the four experts provided forecasts for all of them. The mean of the four forecasts was then taken to represent the expert-elicited forecast for each Phase 9 configuration.

Table 7-4: Sample of 18 possible configurations of Phase 9

$xSLRU$	$xSDP$	xNU	xNM	$OpDem$ (ratio of a day)	Env = OK (prob)
3	3	2	2	4/5	0.3
3	3	3	3	4/5	0.5
4	5	3	2	4/5	0.7
8	8	3	2	4/5	0.5
4	5	4	2	4/5	0.5
3	3	4	2	5/5	0.3
3	3	3	2	5/5	0.5
8	8	4	2	4/5	0.3
4	6	2	3	4/5	0.5

$xSLRU$	$xSDP$	xNU	xNM	$OpDem$ (ratio of a day)	Env = OK (prob)
3	3	4	2	5/5	0.7
4	5	2	2	4/5	0.3
4	6	4	3	5/5	0.7
8	8	3	3	5/5	0.7
4	6	3	3	5/5	0.5
8	8	4	3	5/5	0.7
4	5	4	2	5/5	0.5
4	5	2	2	5/5	0.5
4	5	3	2	5/5	0.3

Results from these 18 forecasts are shown over Figure 7-6 and Figure 7-7. In each figure, the same set of 18 boxplots have been reproduced to show the distribution of the Phase 9 number of failures across the 100 simulation replications for each of the 18 configurations. The boxes in each case include the inter-quartile range of the number of failures from the 100 replications. The crosses indicate outlying values in the simulation results.

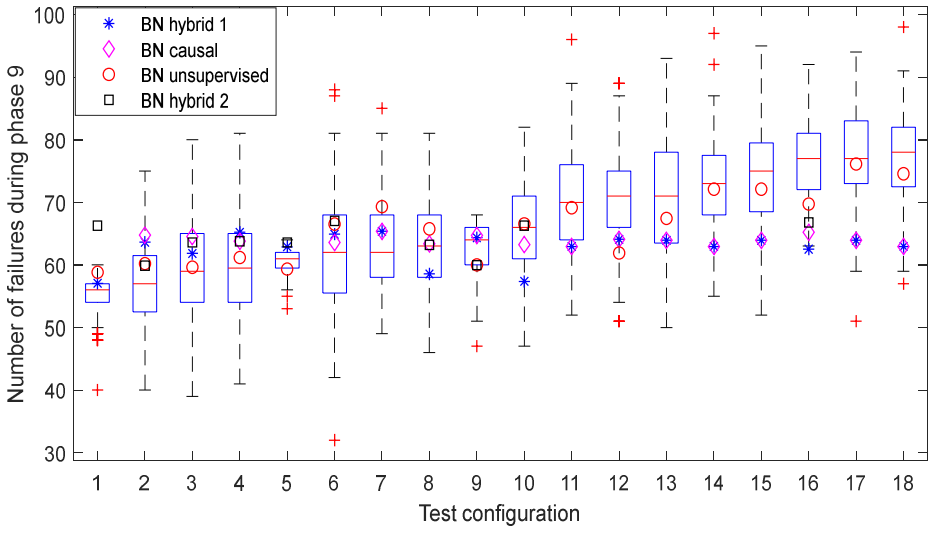


Figure 7-6: A comparison of the BN models' forecasts and the simulation results

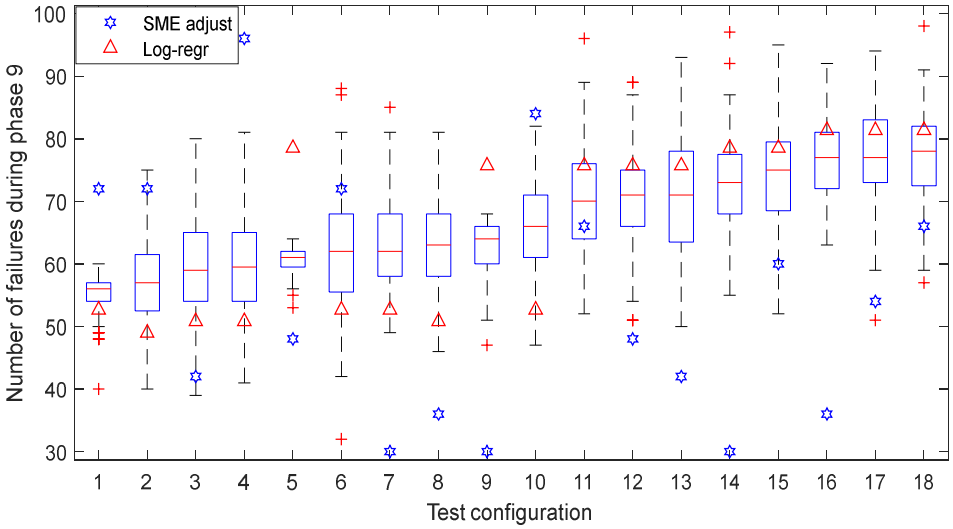


Figure 7-7: A comparison of the regression and the mean SME forecasts and the simulation results

Overlaid on each boxplot are the forecasts for that Phase 9 configuration. In Figure 7-6, forecasts from each of the four BN models are displayed (BN hybrid 1, BN hybrid 2, BN causal, BN unsupervised) in addition to the boxplots of the simulation results. In Figure 7-7, the logistic regression and expert-adjusted forecasts are given (Log-regr, SME adjust) in addition to the boxplots. The vertical axes of these figures record the number of failures for Phase 9, either observed from the Phase 9 simulation results or forecast by one of the considered models. The 18 Phase 9 configurations are arranged in increasing order of the median number of failures obtained from the 100 replications of each of them. Furthermore, Figure 7-8 presents the histograms of each of the sampled configuration in the same order as it is presented in the figures with the boxplots.

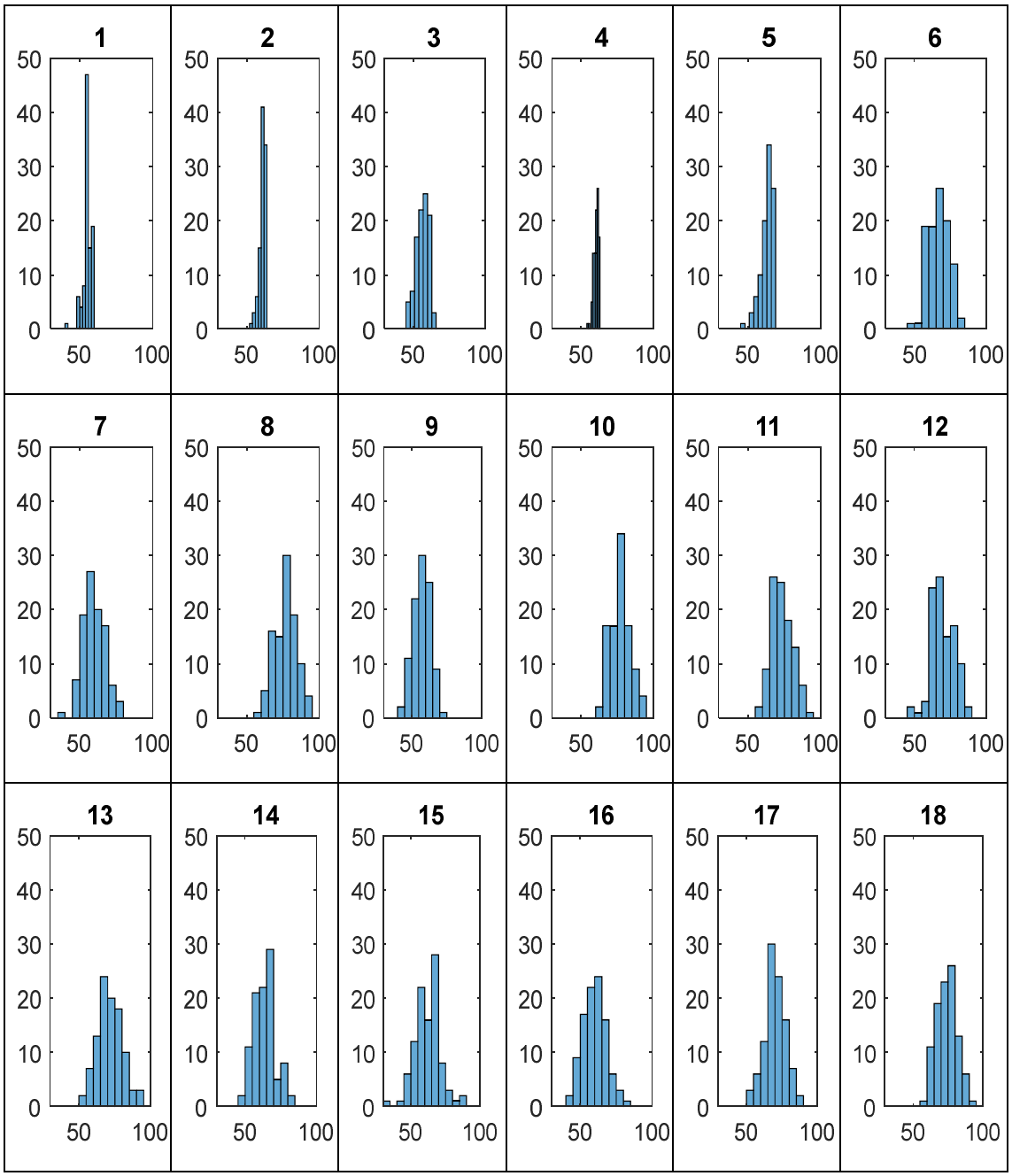


Figure 7-8: Histograms of the sample of 18 configurations of Phase 9

There are two observations that are important to make from Figure 7-6, Figure 7-7 and Figure 7-8, and which though reflect not only the 18 cases but all 144 cases of Table 7-3:

1. The range of demand values is quite large. This suggests that just by chance there can be an output that is quite distant from a model's forecast

value of location. Moreover, this also supports the use of accuracy implication metrics as well as the accuracy metrics

2. Even though the outputs indicate that the experienced demand could be less than 10 units per month and therefore can be categorised as “low volume Stock Keeping Units (SKUs)” (Fildes et al., 2009), none of the values that were acquired had zero number of failures in any of the months, which makes these outputs non-intermittent. However, as seen in the next scenario, when the more complicated EBS is introduced by the use of other components as well, the scenario did provide intermittent data. As also discussed in Appendix B, this observation demonstrates the importance of considering the interactions among the components’ failures, and, as shown, BNs can be a very enlightening tool in this perspective. What is shown in the Appendix is that the BNs’ DAGs provide a visual representation of the associations among the variables, and in this way the user can identify which factors could have a driving effect on the number of failures experienced

The discussion now moves on to using the accuracy metrics and the accuracy implication metrics in order to evaluate the performance of the forecast models.

7.2.4 Forecast Models’ Evaluation

7.2.4.1 Accuracy Metrics

Firstly the accuracy metrics were calculated for all the forecast models using just the 18 cases. This was done because it was not practical to ask the SMEs to provide judgementally adjusted SES forecasts for all the 144 alternative futures that were examined. Then the accuracy metrics were calculated for the four different BN approaches and the logistic regression models using all 144.

As discussed in Section 5.2.3.3, the $MASE$ metric was estimated as the arithmetic mean for the 100 replications of each alternative and then the resulting summary was the geometric mean for the 18 and for the 144 arithmetic means.

The results from the 18 cases (replicated 100 times each) were as follows:

Table 7-5: MASE outputs using just 18 of the 144 Phase 9 alternatives[24]

SN	Models	$MASE$
1	Unsupervised learning BN	1.7336
2	Hybrid BN that maintained the elicited structure and then added machine learning (hybrid 1)	1.7727
3	Hybrid BN that used the elicited structure as a starting DAG for machine learning (hybrid 2)	2.0775
4	Logistic regression	2.3752
5	BN with the elicited DAG	2.5369
6	SMEs' adjusted SES forecasts	4.4427

The results from all the 144 cases, including the above 18 were as follows:

Table 7-6: MASE outputs using all 144 Phase 9 alternatives[25]

SN	Models	$MASE$
1	Unsupervised learning BN	1.6125
2	BN with the elicited DAG	2.0065
3	Hybrid BN that maintained the elicited structure and then added machine learning (hybrid 1)	2.0760
4	Hybrid BN that used the elicited structure as a starting DAG for machine learning (hybrid 2)	2.1240
5	Logistic regression	2.3216

The above outputs show that the unsupervised BN model performed better than the rest, while the SME-adjusted approach was the worst.

7.2.4.2 Accuracy Implication Metrics

Since the SME-adjusted forecast performed a lot worse than the rest in its MASE outputs, the decision was taken not proceed with further consideration of the 18 cases. So, the accuracy implication metrics that were calculated were for the 144 cases with the BNs and the regression models only.

[24] The results are from the lower to the higher MASE output
[25] The results are from the lower to the higher MASE output

A plot of the Holding Volume of inventory (horizontal axis) vs the Operational Availability (vertical axis) for four different target/planned service levels: 80%, 90%, 95% and 99% was built. Due to the convenience of the calculations and its common use among practitioners (Cohen, Zheng and Agrawal, 1997), the service level applied here is the S2 "fill rate" defined as the fraction of the demand that can be satisfied immediately from the on-hand stock (Axsater, 2006).

Following, the assumption of a normal model for the distribution of the demand was taken, which is commonly applied in the literature (Kourentzes, 2013; Syntetos, Boylan and Croston, 2005). In order to get the variance of the forecast errors, the mean squared error (MSE) of each forecast for the monthly realised number of failures during the phases 1 to 8 was estimated and multiplied with the 6 months forecast horizon of Phase 9.

Furthermore, in order to facilitate the explanation of the outputs, Table 7-7 and Table 7-8 have been included. Table 7-7 has each model's root mean squared error which has been used as the standard deviation parameter in the normal distribution model of the demand. Using the standard deviations, in combination with the forecast locations, one can understand whether the 100 replications were correctly included in the calculations. On the other hand, Table 7-8 has the mean signed error of each model as an indicator of the bias (Hoover, 2006).

7.2.4.3 How to Read the Accuracy Implication Metrics

A way to read these types of graphs is to see the relative value of Holding Volume for a given service level, i.e. start from the horizontal axis, and get the curve's projection on the vertical axis. Another way, again for a given service level, is to start with a given Operational Availability, i.e. start from the vertical axis and get the curve's projection on the horizontal axis. In either way, the model that for the same Holding Volume gives higher Operational Availability is better, or the model that gives the same Operational Availability but for a lower relative average Holding Volume is better. In simple terms, the comparison is easy when the service level points on the curves under examination form a parallel line either to the vertical axis, and thus see that they have the same Holding Volume, so it is feasible to compare them on their Operational Availability, or form a parallel to

the horizontal axis, and thus see that they have the same Operational Availability output in order to compare them on the Holding Volume.

The comparison is even easier when the service level points form a line that is more than 90º with the horizontal line. In such a case, the point on the top-left, for the same target service level gives a higher Operational Availability for lower average Holding Volume.

Things get challenging when the previously mentioned line forms an angle that is less than 90º with the horizontal line. In such a case, the model at the top-right corner gives higher Operational Availability but at a higher holding volume. Nevertheless, in such a case the decision maker then needs to choose how much more Operational Availability is acquired when the extra holding volume is spent. Such a decision can be further advised by the other pair of accuracy metrics (on the end of the phase), as well as by the probability of no stock-out service level.

7.2.4.4 Relative Average Holding Volume vs Average Operational Availability

In order to increase the readability of the plot, each value was scaled to the respective worse performance of one of the models for the horizontal axis of the spares Holding Volume. Another reason for scaling is that the plotted data have come from a simulation of a made-up scenario and thus the absolute holding-volume values would mask the relative performance, while they would not give any substantive further information about the importance of the forecasts' differences.

The model to scale upon was chosen to be the logistic regression. So, all the Holding Volumes ("average" and "end") were divided by the highest service level (99%) of the logistic regression respective value. This means that the horizontal axis values for a given service level correspond to each forecast model's Holding Volume performance for that service level relative to how bad a service would be acquired if the logistic regression forecast with the highest planned (most expensive) service level of 99% was used.

The resulting outputs are presented in the following plot (Figure 7-9):

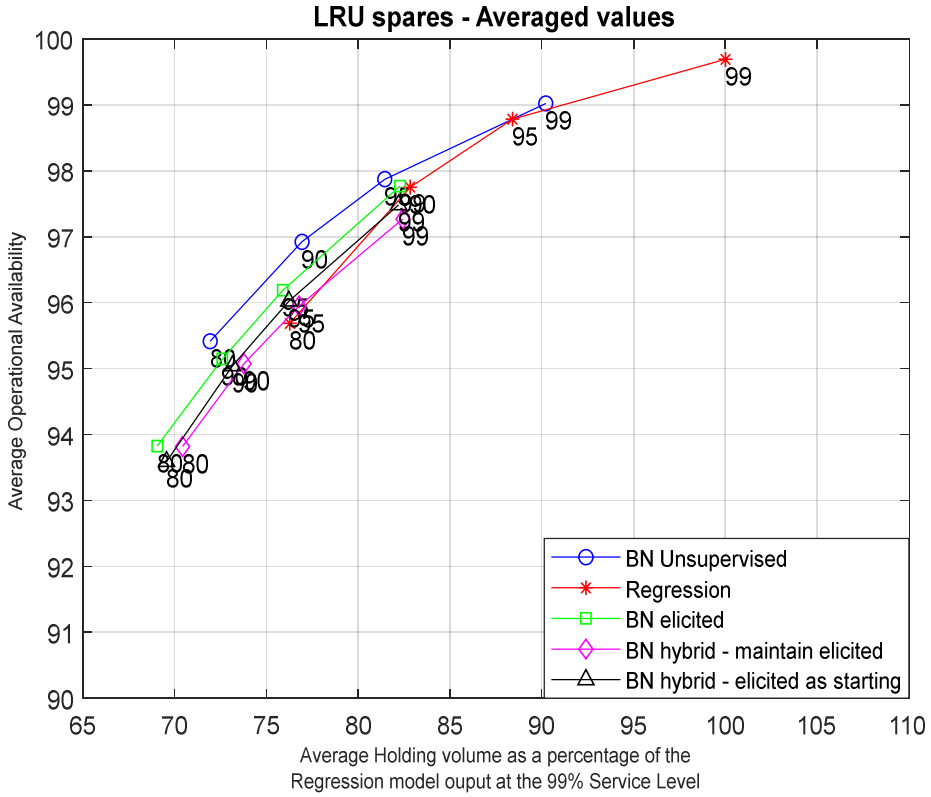


Figure 7-9: Relative average Holding Volumes vs the average Operational Availability [26]

The derived performance is similar but not exactly the same as the MASE accuracy metric output. Nevertheless, these metrics are practical for the decision maker.

The two hybrid structured BNs along with the elicited one have curves that are almost parallel, with the two hybrid almost on top of each other. The parallel positioning makes the comparison convenient since it is easy to see that for any of the four given service levels, for the same relative average Holding Volume the elicited BN performs better/gives higher Operational Availability than the one that used the elicited as a starting graph and this on its turn is (slightly) better than the BN that maintained the elicited structure. This output though is not completely consistent with the output from the MASE accuracy metric comparison, where the order of the last two is the reverse.

[26] The curves are line- interpolating the points for 80%, 90% , 95% and 99% service levels

The elicited BN along with the unsupervised BN and also with the Logistics Regression model fall in the challenging comparison category. For any given service level, the respective points form an approximate line that has a slope less than 90º with the horizontal axis. This means that for any given target service level, the regression model gives higher Operational Availability but for a higher relative average Holding Volume, followed by the unsupervised BN.

This behaviour can be explained by considering the outputs of Table 7-7 and Table 7-8. The unsupervised BN and the regression tend to underforecast less than the other three but also have higher root squared errors and therefore they used higher standard deviations in their (normal) demand distribution models. Hence, for any given target service level the suggestion of the respective demand forecast model was affected by its bias but mostly by its variance. Consequently, for the case of the unsupervised BN and for the regression model, they tended to suggest higher values for spares to be kept.

Table 7-7: Root squared errors of the models

Model	RSE[27] for LRU
BN Unsupervised	7.02
Regression	8.87
BN elicited	5.47
BN hybrid – maintained elicited	4.97
BN hybrid – elicited as starting	5.23

[27] Rounded to the second decimal place

Table 7-8: Mean Signed Error (as an indicator of bias) of the models[28]

Model	Mean Signed Error[29] for LRU
BN Unsupervised	+0.12
Regression	+0.81
BN elicited	+2.66
BN hybrid – maintained elicited	+1.49
BN hybrid – elicited as starting	+2.12

7.2.4.5 Holding Volume vs Operational Availability at the end of the final phase

The results from these plots are similar to the ones with the average values:

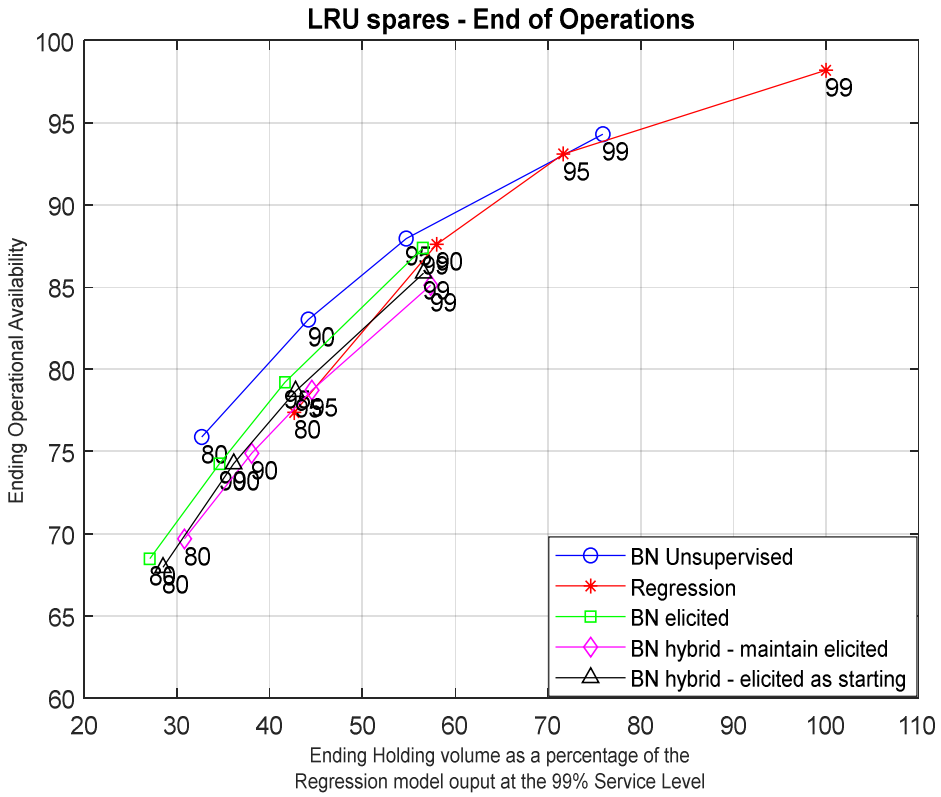


Figure 7-10: Relative Holding Volumes at the end of the final phase vs the Operational Availability at the end of the phase

An interesting observation is that the plotting scales of both the horizontal and the vertical axes of Figure 7-9 and Figure 7-10 are different. The values were

[28] The positive sign indicates underforecasting
[29] Rounded to the second decimal

acquired from averaging (Figure 7-9) are less spread than the ones that acquired from the end of the Operations (Figure 7-10). This is an observation that is also seen in the second, more elaborate scenario (Section 7.3). The reason for this observation is because at the end of the Operations there is a tendency to find low values of Operational Availability and of Holding Volumes more often than when the average outputs of the whole period are used. However, this also stresses the importance of incorporating both measures in order to inform decisions as discussed in Sections 5.2.5 and 5.2.6.

7.2.4.6 Probability of no Stock-Outs During the Whole Phase

The results of this metric are presented in the following Table 7-9:

Table[30] 7-9: Probability of no stock-outs during the whole phase, given the four different fill-rates

		BN Unsupervised	Regression	BN elicited	BN hybrid - maintain elicited	BN hybrid - elicited as starting
Fill-rate	80%	***94.18***	**94.40**	92.39	92.46	92.24
	90%	***96.00***	**96.91**	93.93	93.94	93.84
	95%	***97.16***	**98.26**	95.12	94.94	94.97
	99%	***98.60***	**99.51**	97.02	96.48	96.70

This output shows that the regression model as used in the Stock Management System provides higher probability of no stock-outs than the other models during the whole period of the final phase, followed by the unsupervised BN model. This is an implication of interest to the mechanic as a customer of the inventory. However, when seen in combination with the figures earlier, this higher effectiveness comes at the cost of more inventory volume held both throughout

[30] The outputs that scored higher are in bold, while the second higher are in bold and italics

(average Holding Volume: x-axis of Figure 7-9) but especially as an amount of unused leftovers (Holding Volume at the end of the phase: x-axis of Figure 7-10).

Nevertheless, this accuracy implication metric shows that with any of the models one would, on average, not expect to have a lack of inventory service for a long period. The longest period of being without a spare would be (100-92.24)% of 6 months, i.e. about 14 days on average if the BN hybrid model was adopted which used the elicited BN as a starting structure.

7.2.5 Discussion

Considering both the accuracy and the accuracy implication outputs, it can be seen that the unsupervised BN performs better in the first but the same cannot be claimed as clearly for the second. The regression model seems to be providing better customer-oriented outputs (average and at the end of the period Operational Availability for the operations, and probability of no stock-outs for the repair activities of the mechanics), but at the cost of less efficiency both averaged through the period (more holding costs) and when the operations have finished (more "waste"). From a modelling perspective this difference between the accuracy and the accuracy implication metrics is mainly due to two reasons:

Firstly, the regression model gave higher variance/standard deviation than the rest. This fact also highlights the importance of this parameter in the implications (Boylan and Syntetos, 2006; Willemain, 2006; Willemain, Smart, and Schwarz, 2004).

Secondly, even though it is very often used in practice (Kourentzes 2013, p.203; Syntetos et al. 2009, p.72), the normal distribution model for the number of demands might not be the most appropriate. This observations has resulted in hindsight coming from plots like the ones in Figure 7-8 but even more from what is presented next from the second scenario with the more elaborate EBS (Section 7.3). This observation is discussed more fully in Appendix B.

The combination of the two reasons above suggests that when the four target service levels were set and used the normal distribution as defined by each model's forecast mean and resulting standard error, resulted in certain quantiles

that corresponded to the chosen service levels. These normal model's quantiles were compared with the actual values from the simulation of the 144 times 100 different outputs of Phase 9. The models' suggested values were affected mostly by the size of the standard error which resulted in suggesting/forecasting higher values than actually experienced and thus, high Holding Volumes.

7.3 Simulation Support Scenario – Case 2

In this scenario, the same Operational and Support Chain concept was used but the complexity of the UAV's Equipment Breakdown Structure (EBS) was increased. As before, each UAV was composed of a single LRU that could be repaired at the CENTRAL depot by the replacement of a single Disposable Part (DP) kept in the same store as the LRUs. Furthermore, a UAV also had a Partially Repairable Unit (PRU) which could also be repaired – without any additional parts needed – at the CENTRAL depot. However, a PRU could be diagnosed as "beyond repair" and thus be discarded. In such a case an order was placed at the manufacturer. Just like for the DP, due to the assumed high cost of a PRU, the depot was using an (S, S-1) inventory policy and thus initiated a resupply order whenever there was a single PRU unit removed from the PRU inventory, while a repair activity was also initiated immediately when one was needed. Additionally, a UAV also had a Discardable Unit (DU) as a component of its structure. A DU could not be repaired, so every time one was removed from the DU inventory in order to repair a UAV at the FORWARD support in the first-line, due to the again (S, S-1) inventory policy another unit was ordered from the manufacturer. Finally, once more the case where the systems' innate failure rates would change with age was not considered.

7.3.1 Scenario for Dataset Generation

The chosen scenario involved a single iteration of the following consecutive eight phases (Table 7-10):

Table 7-10: Scenario Phases

Phase	Duration (Months)	$xSLRU$	$xSDP$	$xSPRU$	$xSDU$	xNU	xNM	$xNTr$	$OpDem$ (ratio of a day)	Env = OK (prob)
1	1	3	3	3	3	2	2	1	4/5	0.6
2	1	3	3	3	4	3	2	2	4/5	0.6
3	4	3	3	4	4	4	3	3	4/5	0.6
4	6	4	5	5	6	4	3	3	4/5	0.6
5	4	4	6	6	6	4	3	3	5/5	0.6
6	4	4	6	4	5	3	2	2	5/5	0.6
7	2	3	5	3	5	3	2	2	5/5	0.6
8	2	3	5	3	4	2	2	1	5/5	0.6

Again there were eight phases before the final, and the assumed story regarding the operational demand ($OpDem$) and the environmental conditions as shown in Table 7-10 above, was similar to the one in the previous scenario in Case 1.

7.3.2 Simulation of Test Data to Allow Forecast Comparison

As before, the end of Phase 8 provided the initial conditions for a follow-on ninth phase of six months' duration that was used to evaluate the performance of the demand prediction models.

The following table summarises the full factorial experimental design of the contexts which were simulated for 100 repetitions each in order to provide the data needed to evaluate the forecast models. The design produced 512 different contexts:

Table 7-11: The combinations of Phase 9 configurations that constituted the test dataset

	$xSLRU$	$xSDP$	$xSPRU$	$xSDU$	xNU	xNM	$xNTr$	$OpDem$ (ratio of a day)	Env = OK (prob)
Phase 9	3 , 4	3 , 6	3 , 4	4 , 6	2 , 3	2 , 3	2 , 3	4/5 , 5/5	0.3 , 0.7

7.3.3 Forecasting Approaches Employed

Using what can be known in advance, i.e. the values in Table 7-11, the main variables of interest were the probability of experiencing a failure incident in either an LRU, PRU or DU component in any given hour.

Yet again the interest was to examine the same types of demand forecast models, i.e. BNs built through unsupervised learning, BNs that their DAG is elicited from experts, BNs with DAG built in a hybrid way either by maintaining the elicited DAG and building on it through machine learning, or using the elicited as a starting DAG and then using machine learning, and finally a logistic regression. However, another consideration came to play in Case 2 where a more elaborate EBS scenario was used. The question was whether to have a single BN that includes the FRT nodes of all the components or to have an individual BN for each FRT. Consequently, the following demand forecasting models were compared:

Table 7-12: List of the models that have been explored for the modelling of the demand in the second scenario

ID	Models explored
BN1	A BN which used unsupervised learning for its DAG and that had only the FRT of 1. The LRU as an FRT node, or 2. The PRU as an FRT node, or 3. The DU as an FRT node This means that **three** different BN models were built
BN2	A BN that its DAG structure was elicited and that had only the FRT of 1. The LRU as an FRT node, or 2. The PRU as an FRT node, or 3. The DU as an FRT node This means that **three** different BN models were built
BN3	A BN with a hybrid DAG which was developed using the elicited structure as its start and that had only the FRT of 1. The LRU as an FRT node, or 2. The PRU as an FRT node, or 3. The DU as an FRT node This means that **three** different BN models were built (corresponds to BN hybrid 2 of scenario Case 1)

ID	Models explored
BN4	A BN with a hybrid DAG to which the elicited structure was maintained and that had only the FRT of 1. The LRU as an FRT node, or 2. The PRU as an FRT node, or 3. The DU as an FRT node This means that **three** different BN models were built (corresponds to BN hybrid 1 of scenario Case 1)
BN5	A **single** BN which used unsupervised learning for its DAG and that had 1. The FRT of the LRU as one of its nodes 2. The FRT of the PRU as one of its nodes 3. The FRT of the DU as one of its nodes
BN6	A **single** BN that has its DAG structure elicited and that had 1. The FRT of the LRU as one of its nodes 2. The FRT of the PRU as one of its nodes 3. The FRT of the DU as one of its nodes
BN7	A **single** BN with a hybrid DAG which was developed using the elicited structure as its start and that had 1. The FRT of the LRU as one of its nodes 2. The FRT of the PRU as one of its nodes 3. The FRT of the DU as one of its nodes (Corresponds to BN hybrid 2 of scenario Case 1)
BN8	A **single** BN with a hybrid DAG to which the elicited structure was maintained and that had 1. The FRT of the LRU as one of its nodes 2. The FRT of the PRU as one of its nodes 3. The FRT of the DU as one of its nodes (Corresponds to BN hybrid 1 of scenario Case 1)
M9	A logistic regression for the FRT of 1. The LRU, or 2. The PRU, or 3. The DU This means that **three** different regression models were built

In summary, what was developed consisted of eight BN modelling approaches and a logistic regression for demand forecast of each of the three components. Therefore, the number of models that were built in total were 19: three BN1, three BN2, three BN3, three BN4, one BN5, one BN6, one BN7, one BN8 and three M9.

All the BN models were built in R, using the bnlearn package (Nagarajan, Scutari and Lebre, 2013). The DAGs of the BNs and the coefficients of the regression models are presented in the Appendix A.

The discussion now moves on to using the accuracy metrics and the accuracy implication metrics in order to evaluate the performance of the forecast models.

7.3.4 Forecast Models' Evaluation

7.3.4.1 Accuracy Metrics

The $MASE$ metric was again calculated as the arithmetic mean for the 100 replications of each alternative and then the geometric mean for the 512 arithmetic means respectively. Furthermore, this process was performed once per individual component and once overall:

Table 7-13: MASE values for the 512 cases for the forecast of the LRU/DP only

SN	ID	Models	$MASE$
1	BN1	Unsupervised learning BN – a different model for every component	1.9719
2	BN5	Unsupervised learning BN – all components in a single model	2.2005
3	BN4	Hybrid BN that maintained the elicited structure and then added machine learning - a different model for every component (corresponds to BN hybrid 1 of scenario Case 1)	2.7106
4	BN3	Hybrid BN that used the elicited structure as a starting DAG for machine learning - a different model for every component (corresponds to BN hybrid 2 of scenario Case 1)	2.7106
5	BN2	BN with the elicited DAG - a different model for every component	2.8862
6	M9	Logistic regression	3.5030
7	BN7	Hybrid BN that used the elicited structure as a starting DAG for machine learning - all components in a single model (corresponds to BN hybrid 2 of scenario Case 1)	10.4920
8	BN6	BN with the elicited DAG - all components in a single model	10.6798
9	BN8	Hybrid BN that maintained the elicited structure and then added machine learning - all components in a single model (corresponds to BN hybrid 1 of scenario Case 1)	13.2647

Table 7-14: MASE values for the 512 cases for the forecast of the PRU only

SN	ID	Models	$MASE$
1	BN1	Unsupervised learning BN – a different model for every component	2.4680
2	BN2	BN with the elicited DAG - a different model for every component	2.6555
3	BN4	Hybrid BN that maintained the elicited structure and then added machine learning - a different model for every component (corresponds to BN hybrid 1 of scenario Case 1)	2.9651
4	BN5	Unsupervised learning BN – all components in a single model	3.0390
5	BN3	Hybrid BN that used the elicited structure as a starting DAG for machine learning - a different model for every component (corresponds to BN hybrid 2 of scenario Case 1)	3.5019
6	M9	Logistic regression	4.1482
7	BN7	Hybrid BN that used the elicited structure as a starting DAG for machine learning – all components in a single model (corresponds to BN hybrid 2 of scenario Case 1)	9.8553
8	BN6	BN with the elicited DAG – all components in a single model	10.3616
9	BN8	Hybrid BN that maintained the elicited structure and then added machine learning – all components in a single model (corresponds to BN hybrid 1 of scenario Case 1)	11.5068

Table 7-15: MASE values for the 512 cases for the forecast of the DU only

SN	ID	Models	$MASE$
1	BN1	Unsupervised learning BN – a different model for every component	2.8817
2	BN4	Hybrid BN that maintained the elicited structure and then added machine learning - a different model for every component (corresponds to BN hybrid 1 of scenario Case 1)	3.2480
3	BN2	BN with the elicited DAG - a different model for every component	3.2919
4	M9	Logistic regression	3.5570
5	BN5	Unsupervised learning BN – all components in a single model	3.6811

SN	ID	Models	$MASE$
6	BN3	Hybrid BN that used the elicited structure as a starting DAG for machine learning - a different model for every component (corresponds to BN hybrid 2 of scenario Case 1)	3.9711
7	BN7	Hybrid BN that used the elicited structure as a starting DAG for machine learning – all components in a single model (corresponds to BN hybrid 2 of scenario Case 1)	11.5251
8	BN6	BN with the elicited DAG – all components in a single model	12.1332
9	BN8	Hybrid BN that maintained the elicited structure and then added machine learning – all components in a single model (corresponds to BN hybrid 1 of scenario Case 1)	13.4399

Table 7-16: MASE values for the 512 cases for the forecast of All parts

SN	ID	Models	$MASE$
1	BN1	Unsupervised learning BN – a different model for every component	2.4115
2	BN5	Unsupervised learning BN – all components in a single model	2.9090
3	BN2	BN with the elicited DAG - a different model for every component	2.9330
4	BN4	Hybrid BN that maintained the elicited structure and then added machine learning - a different model for every component (corresponds to BN hybrid 1 of scenario Case 1)	2.9665
5	BN3	Hybrid BN that used the elicited structure as a starting DAG for machine learning - a different model for every component (corresponds to BN hybrid 2 of scenario Case 1)	3.3530
6	M9	Logistic regression	3.7250
7	BN7	Hybrid BN that used the elicited structure as a starting DAG for machine learning – all components in a single model (corresponds to BN hybrid 2 of scenario Case 1)	10.6021
8	BN6	BN with the elicited DAG – all components in a single model	11.0320
9	BN8	Hybrid BN that maintained the elicited structure and then added machine learning – all components in a single model (corresponds to BN hybrid 1 of scenario Case 1)	12.7062

The above outputs show that the BN models that incorporated all of the components in a single model and used either the elicited DAG or the hybrid DAGs (BNs 6 to 8) performed worse in each comparison. The model that

performed best in all cases was the BN that used unsupervised learning to build DAGs that included only one of the targeted components each time (BN 1).

The performance of the rest of the models varied. The BN that used machine learning but by having all components in a single model (BN 5) was the second best in the MASE comparison using all the parts (Table 7-16). This output is encouraging for practical purposes because it could mean that fewer BN models (a single in the specific case) would need to be built, with very little SMEs engagement, and thus a higher efficiency in the process could be acquired due to the automation. However, this promising performance of BN 5 was not sustained in the individual parts, except for the LRU/DP (Table 7-13). BN 5 had the 4th best performance out of the 9 models for the PRU (Table 7-14), and the 5th best for the DU (Table 7-15).

BN 2, the model that was elicited individually for each part was the 3rd best in the overall parts comparison, which again indicates the effectiveness of careful consideration of the mechanisms that exist among the variables in the demand context. However, again BN 2 was 5th for the LRU/DP, 2nd for the PRU and 3rd for the DU.

The extension of BN 2, the hybrid BN 4 that maintained the elicited structure(s) and then added machine learning, was 4th overall. This output seems to be related to that above regarding the merits of using the understanding of the mechanisms within the demand context, but it also shows that – at least in the examined cases – the extra effort through machine learning did not give better results. BN 4 was 3rd for the LRU/DP and the PRU, while it was 2nd for the DU.

Finally, BN 3, the hybrid that in an effort to replace random starting of the conventional machine learning process of BN 1 (and BN 5) started from the assumed knowledge of the SMEs, was 5th in the overall comparison. The observations are similar to the ones above, i.e. that the random starting seemed better. BN 3 was 4th for the LRU/DP, 5th for the PRU and 6th for the DU.

On the other hand, the logistic regression was 6th for the overall parts comparison, for the LRU/DP, and for the PRU, while it was 4th for the DU.

7.3.4.2 Accuracy Implication Metrics

Again, the (relative) Holding Volume of inventory (horizontal axis) versus the Operational Availability (vertical axis) plots of four different target/planned service levels were developed: 80%, 90%, 95% and 99%. Furthermore, in order to study the outputs in more detail, the forecast implications were included both for the three individual components and for their overall summary. Additionally, given the large difference in the MASE accuracy of the BN models BN 6, 7 and 8 (Table 7-13, Table 7-14, Table 7-15 and Table 7-16), and in order to reduce the cluster, these models were placed in different plots and then compared to the regression model.

The regression model has been included in both the comparisons of BN 1 to 5 and of BN 6 to 8. Moreover, the two different groups have different in the horizontal and vertical axes scales due to the distance in their respective range of values, especially the Holding Volume.

Finally, to support further explanation of the outputs, Table 7-17 and Table 7-18 were produced. Table 7-17 includes each model's root mean squared errors for each component which was used as the standard deviation parameter in the normal distribution models of the demand, and Table 7-18 presents the mean signed error of each model as an indicator of the bias (Hoover, 2006).

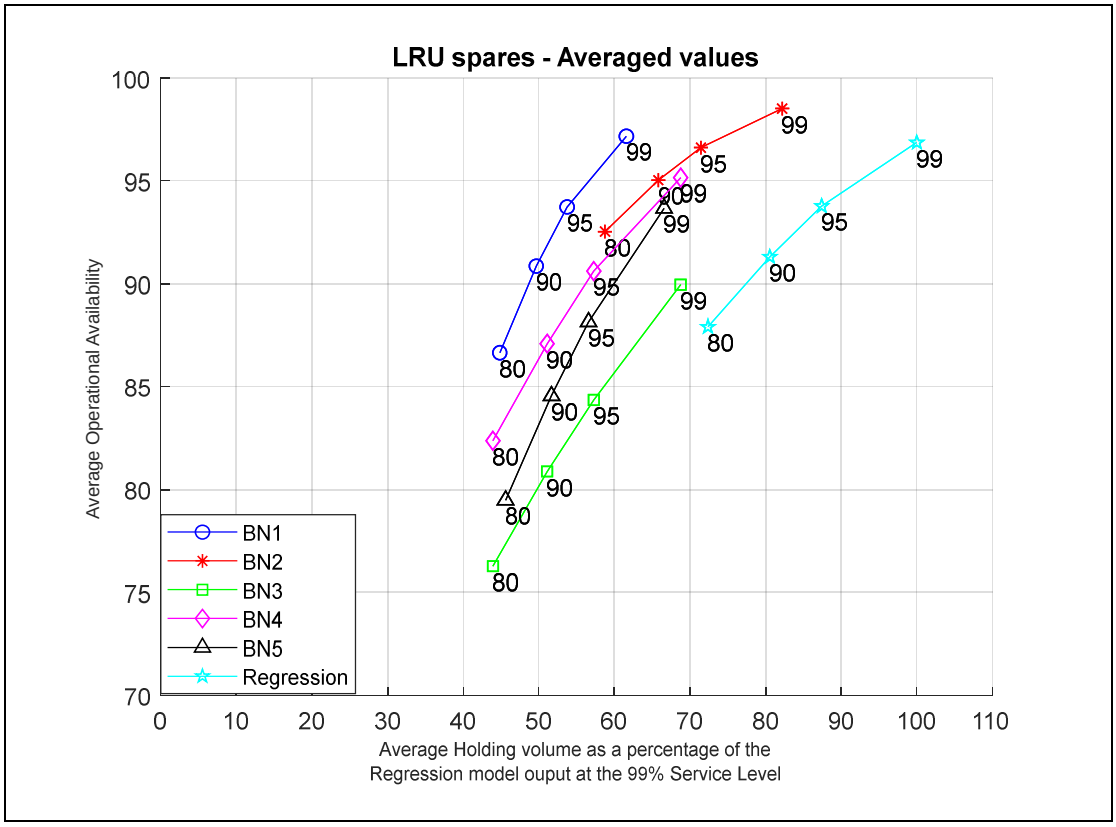
7.3.4.3 Relative Average Holding Volume vs Average Operational Availability

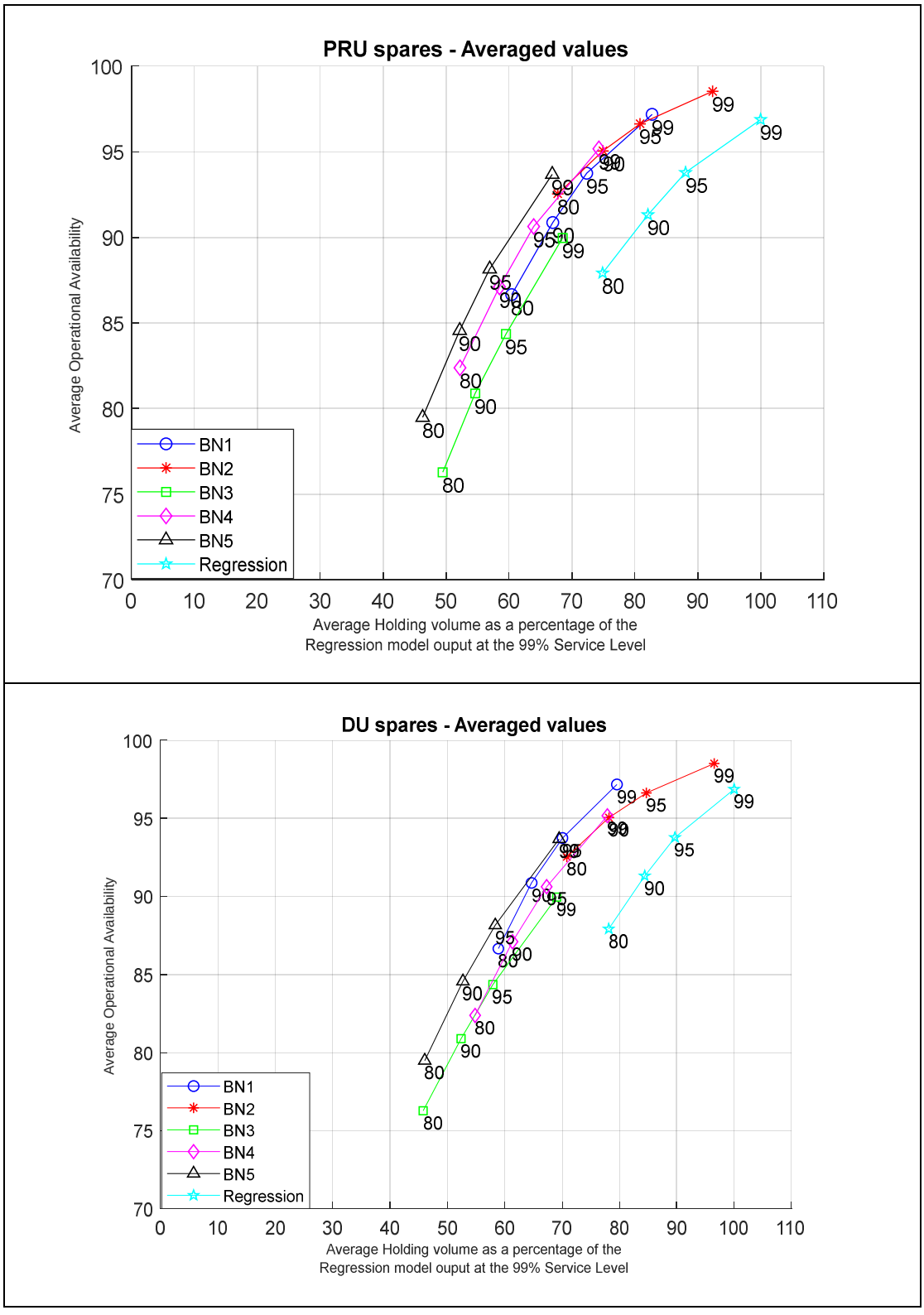
Discussion now starts with the plots in Figure 7-12. The poor performance of the BN models 7 and 8 is obvious for all components and of course overall as well. For high holding costs, their suggested forecasts cannot deliver average Operational Availability beyond 65% even at the 99% targeted service level.

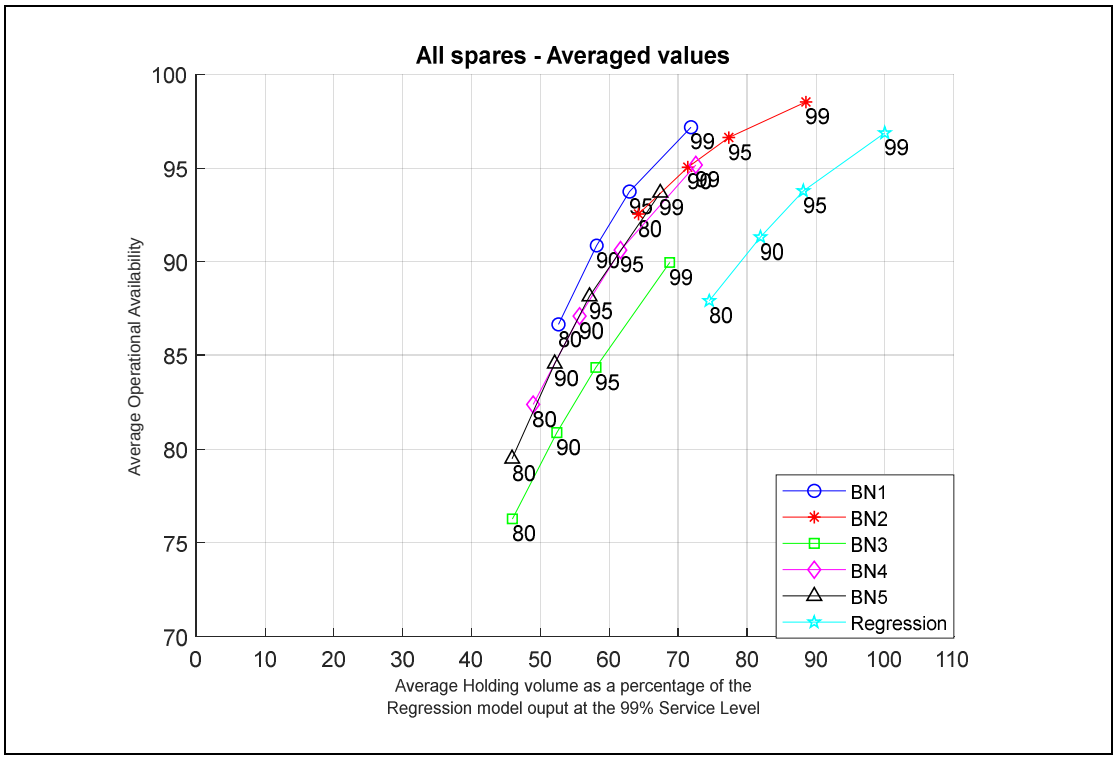
At first glance, this result may look counter-intuitive. How can the model result in such relatively high Holding Volumes and at the same time give such low Operational Availability in every single component and overall? The reason is that the forecast models BN 7 and 8, in some of the cases – but not in all – have large negative errors (i.e. the forecasts are a lot higher than the actual values; see also Table 7-18). The resulting high inventory volume does not produce equally high

levels of Operational Availability, since the latter cannot go beyond 100% of the supported units and for the cases of these BN models – as the plot shows - it was usually low. Consequently, the mean values of the Holding Volumes are affected a great deal by such outlying outputs, while the Operational Availability means are not.

The case of BN 6 is similar but not exactly the same. The model's performance is giving good Operational Availability on average, but at the expense of high relative average Holding Volumes. The behaviour of the BNs 6, 7 and 8 can be partly explained by their comparatively high mean signed errors (Table 7-18).

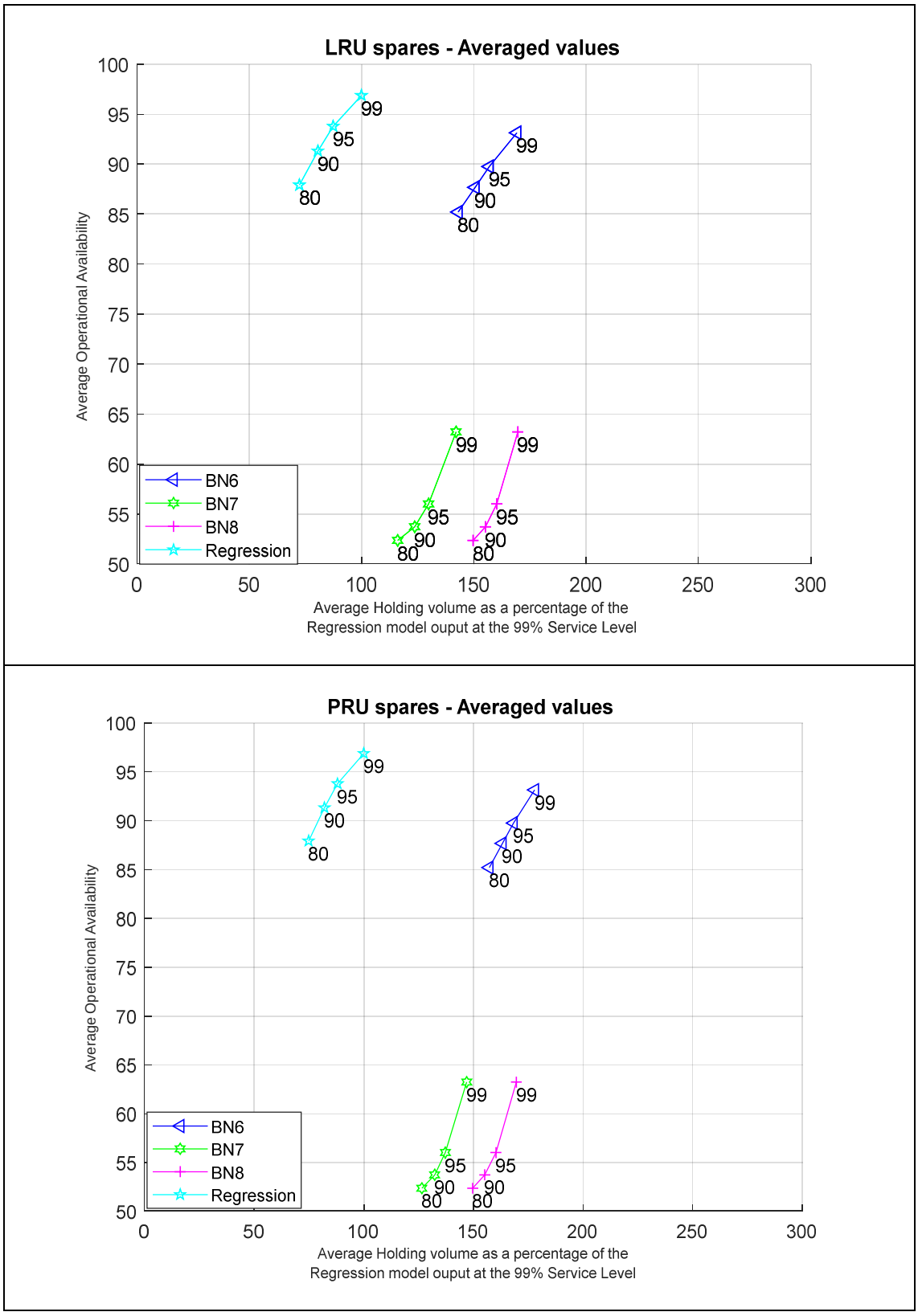


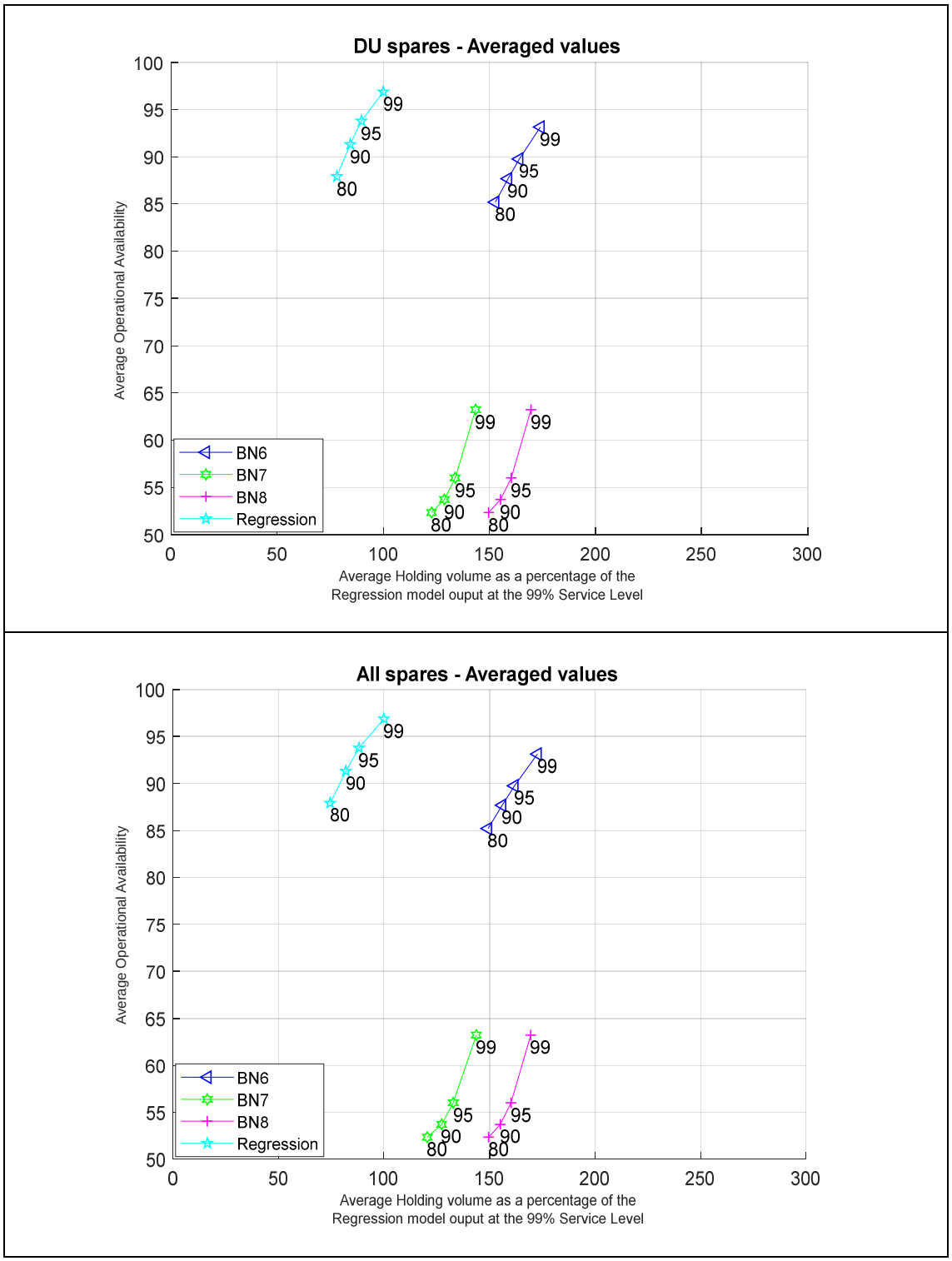

PRU spares - Averaged values
Average Operational Availability
Average Holding volume as a percentage of the
Regression model ouput at the 99% Service Level
BN1
BN2
BN3
BN4
BN5
Regression
DU spares - Averaged values
Average Operational Availability
Average Holding volume as a percentage of the
Regression model ouput at the 99% Service Level
BN1
BN2
BN3
BN4
BN5
Regression



Figure[31] 7-11: Relative average Holding Volumes vs the average Operational Availability (four plots)

[31] One plot for each of the forecast spare component (LRU, PRU and DU) and one overall. The present plot presents the BN models 1 to 5 and the logistic regression models as per the list in Table 7-12


LRU spares - Averaged values
Average Operational Availability
Average Holding volume as a percentage of the
Regression model ouput at the 99% Service Level
BN6
BN7
BN8
Regression
PRU spares - Averaged values
Average Operational Availability
Average Holding volume as a percentage of the
Regression model ouput at the 99% Service Level
BN6
BN7
BN8
Regression



Figure[32] 7-12: Relative average Holding Volumes vs the average Operational Availability (four plots)

[32] One plot for each of the forecast spare component (LRU, PRU and DU) and one overall. The present plot presents the BN models 6 to 8 and the logistic regression models as per the list in Table 7-12

Looking at Figure 7-11, BN 1 is consistently parallel and to the left of the regression, which shows that BN 1 for any of the tested target service levels, delivers the same average Operational Availability as the regression but for lower average relative Holding Volume. This output is supported by the MASE accuracy metric. The explanation of this output can be inferred by comparing the two models' outputs in Table 7-17 and Table 7-18. The regression model tends to overforecast – especially in the LRU case – and also has a higher root squared error.

BN 2 performs better than the regression model too. For any given target service level, BN 2 points on the plot are above and to the left of the respective ones of the regression, which shows that it gives higher average Operational Availability for lower average relative Holding Volumes. As before, BN 2 has better mean signed error values and lower root mean squared error as well.

The situation is not as clear in the comparison of the regression and BNs 3, 4 and 5. For any of the target service levels these BNs provide lower average Operational Availability than the regression model but they also use lower average relative Holding Volumes.

The same dilemma comes from the comparison of BNs 3, 4 and 5 to BN 2 which is above and to their right.

When comparing BN 1 to BN 3, 4 and 5 on the LRU component, the first performs clearly better followed by 4, 5 and then 3. However, for the PRU and the DU, BN 3 has not done as well as BNs 4 and 5, while the comparison among the latter two and BN 1 is not as clear. Finally, in the curves with the overall parts, the order is BN 1, 4, 5 and 3.

Similarly, to the Case 1 scenario, there seems to be a slight inconsistency between these results and the MASE accuracy ones, where considering MASE the BN 1 performed better in all comparisons. The reason for this inconsistency is that in MASE the comparison was of the performance of the location parameter of the distribution only, while with the accuracy implication metrics the comparison was of the output of the location as used in the (normal distribution) model along with the respective estimate of the variance.

Table 7-17: Root squared errors of each modelling approach

Model[33]	RSE for LRU	RSE for PRU	RSE for DU
BN1	7.47	5.98	5.17
BN2	10.34	6.38	6.13
BN3	11.23	5.67	6.13
BN4	11.23	6.07	5.79
BN5	9.33	5.79	6.04
BN6	11.77	5.72	5.46
BN7	11.80	5.76	5.47
BN8	11.85	5.57	5.30
M9	12.26	6.82	5.07

Table 7-18: Mean Signed Error (as an indicator of bias) of the models[34]

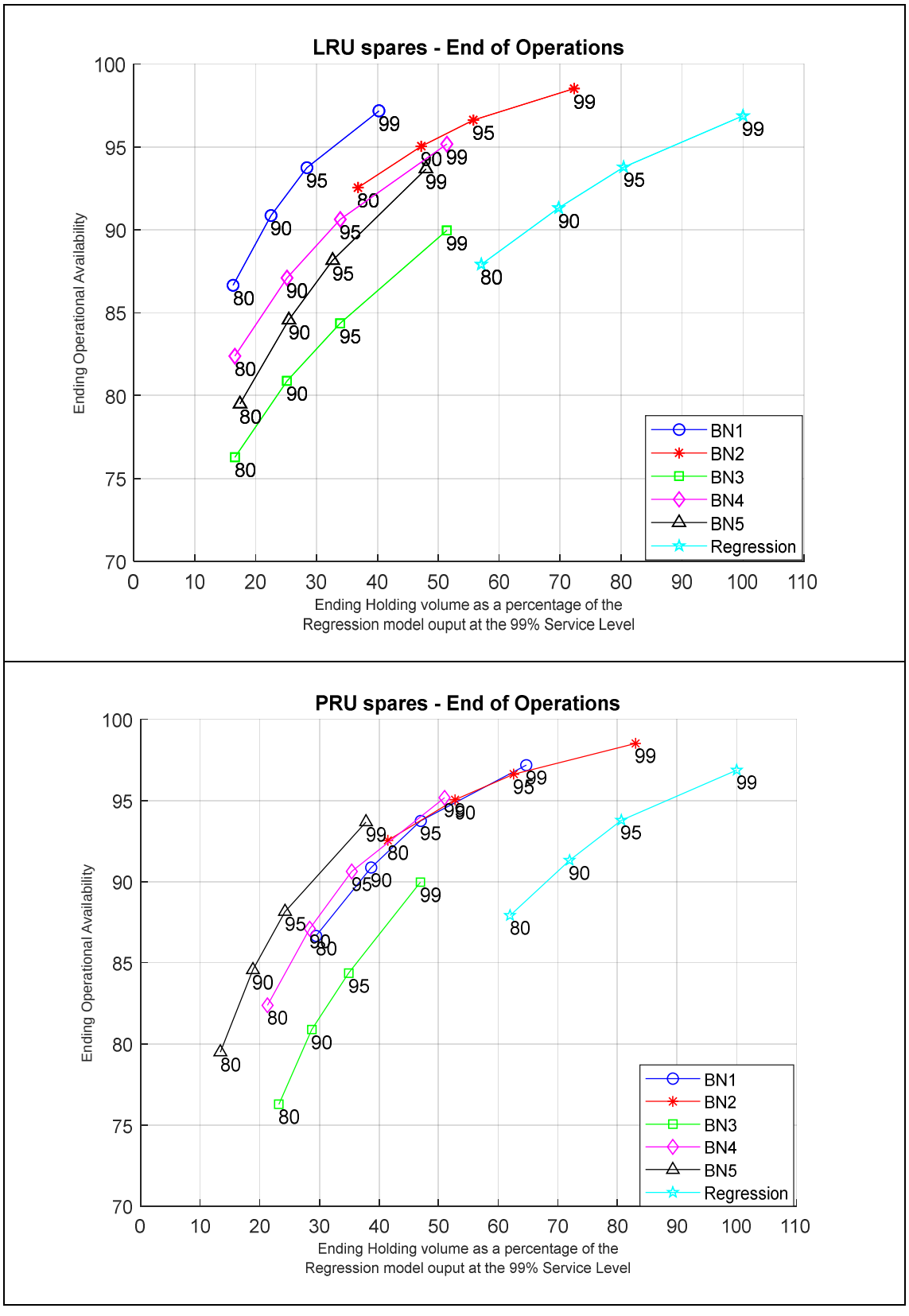
Model[35]	Mean Signed Error for LRU	Mean Signed Error for PRU	Mean Signed Error for DU
BN1	-0.30	+0.93	+0.80
BN2	-7.11	-1.61	-2.85
BN3	+3.74	+6.18	+6.86
BN4	+3.74	+4.53	+2.84
BN5	+0.83	+6.68	+6.50
BN6	-60.92	-35.05	-31.06
BN7	-42.99	-23.14	-20.32
BN8	-56.80	-32.23	-28.54
M9	-14.43	-3.37	-6.39

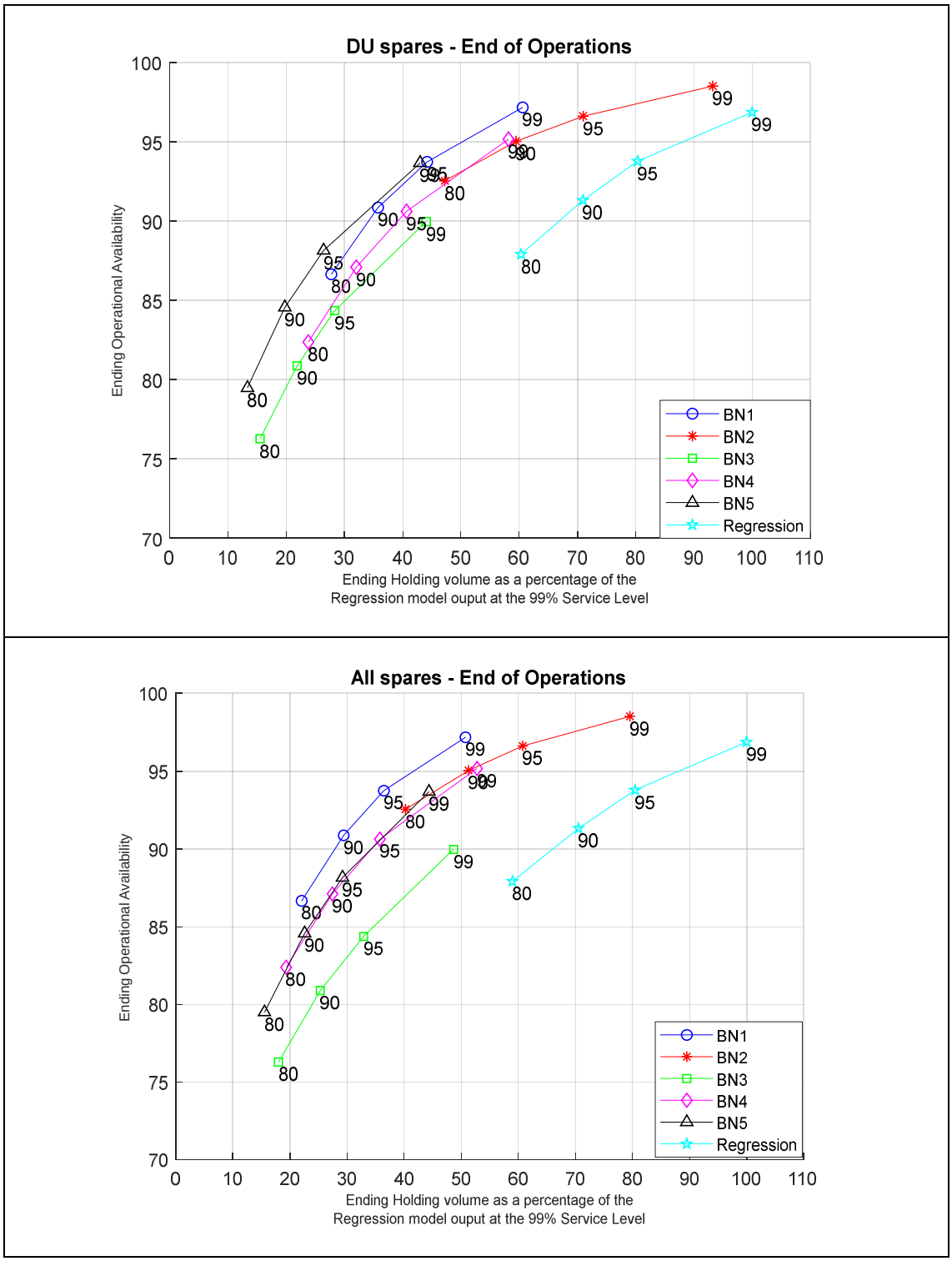
7.3.4.3.1 Holding Volume vs Operational Availability at the End of the Final Phase

The results from the end of the phase plots are similar to the ones that were acquired on average, with the only characteristic difference on the horizontal and

[33] Rounded to the second decimal
[34] The negative sign indicates overforecasting
[35] Rounded to the second decimal

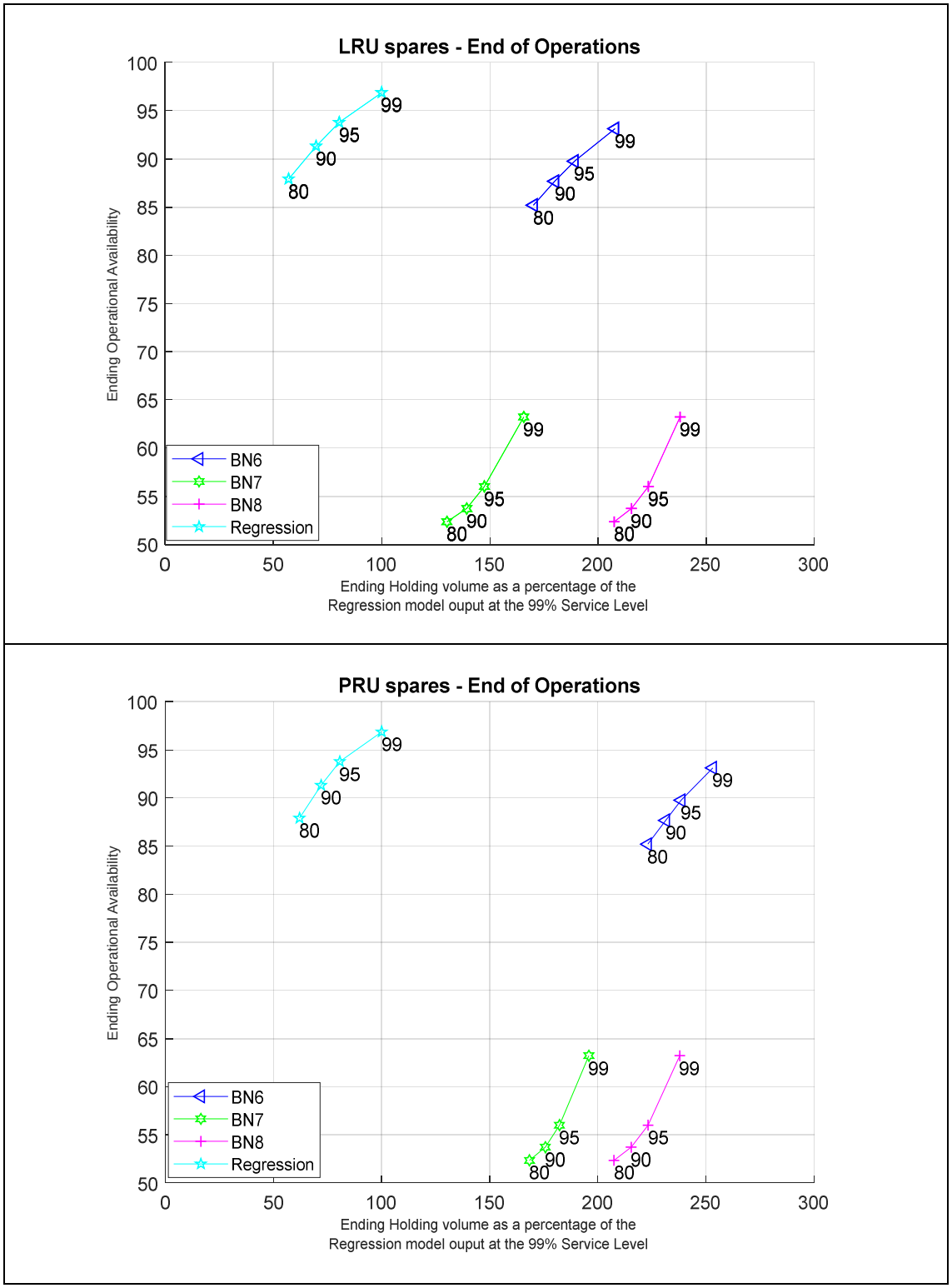
vertical axes scales. As in Case 1 scenario, the plots of the averages have less spread out ranges, which is the effect of averaging.

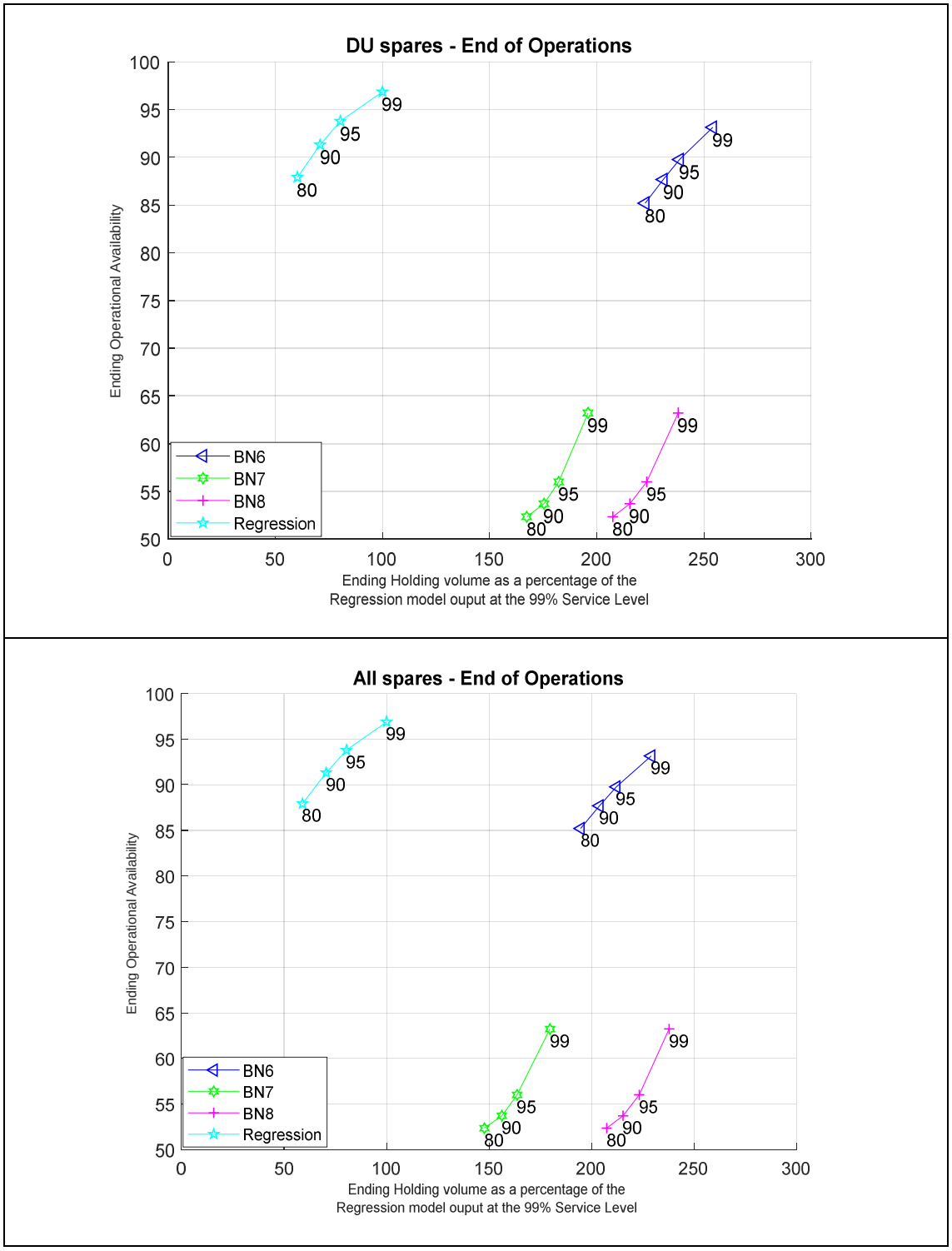




Figure[36] 7-13: Relative Holding Volumes at the end of the final phase vs the Operational Availability at the end of the phase (four plots)

[36] One plot for each of the forecast spare component (LRU, PRU and DU) and one overall. The present plot presents the BN models 1 to 5 and the logistic regression models as per the list in Table 7-12


LRU spares - End of Operations
Ending Operational Availability
Ending Holding volume as a percentage of the
Regression model ouput at the 99% Service Level
BN6
BN7
BN8
Regression
PRU spares - End of Operations
Ending Operational Availability
Ending Holding volume as a percentage of the
Regression model ouput at the 99% Service Level
BN6
BN7
BN8
Regression



Figure[37] 7-14: Relative Holding Volumes at the end of the final phase vs the Operational Availability at the end of the phase (four plots)

[37] One plot for each of the forecast spare component (LRU, PRU and DU) and one overall. The present plot presents the BN models 6 to 8 and the logistic regression models as per the list in Table 7-12

7.3.4.3.2 Probability of no Stock-Outs during the Whole Phase

This output shows that the regression model as used in the Stock Management System provides higher probability of no stock-outs than the other models during the whole period of the final phase, followed by the BN 2 model, in all components and target service levels apart from the PRU. The lowest performance is delivered by BN 3 for an 80% target service level of the PRU component, in which one would expect on average to get a stock-out situation (100-81.78) % of 6 months, i.e. about 33 days before the end of the final phase.

Table 7-19: Probability of no stock-outs during the whole phase, given the four different fill-rates

			BN 1	BN 2	BN 3	BN 4	BN 5	BN 6	BN 7	BN 8	M 9
LRU	Fill-rate	80%	95.67	***98.49***	93.64	93.64	95.70	95.27	92.91	94.32	**99.82**
		90%	98.09	***99.30***	97.39	97.39	98.51	97.72	96.12	96.64	**99.99**
		95%	99.10	***99.65***	99.12	99.12	99.46	99.13	98.12	98.09	**100.00**
		99%	99.79	***99.92***	99.90	99.90	99.91	99.91	99.71	99.52	**100.00**
PRU		80%	***92.55***	**94.65**	81.78	87.39	85.02	88.33	85.44	88.58	89.31
		90%	***94.96***	**96.44**	85.07	90.83	88.99	90.51	88.10	90.57	92.14
		95%	***96.58***	**97.53**	87.45	93.15	91.54	92.32	90.01	91.99	94.21
		99%	***98.46***	**98.85**	91.41	96.40	95.47	95.05	93.37	94.28	96.91
DU		80%	91.70	***95.55***	82.83	89.35	84.65	87.94	84.95	88.37	**96.18**
		90%	94.34	***97.23***	87.23	92.45	89.05	90.26	87.63	90.29	**97.58**
		95%	96.06	***98.22***	90.33	94.75	92.11	92.15	89.64	91.71	**98.51**
		99%	98.14	***99.33***	94.83	97.47	96.21	95.04	93.22	94.42	**99.46**
All		80%	93.31	**96.23**	86.08	90.13	88.46	90.52	87.76	90.43	***95.10***
		90%	95.80	**97.65**	89.90	93.56	92.18	92.83	90.62	92.50	***96.57***
		95%	97.25	**98.47**	92.30	95.67	94.37	94.53	92.59	93.93	***97.57***
		99%	98.80	**99.37**	95.38	97.92	97.20	96.67	95.43	96.07	***98.79***

Another quite interesting observation comes when one looks to identify which of the components gave the highest risk of a stock-out with any of the 9 models. This can be observed by looking at any single individual service level for all the parts, e.g. the 80% row for the LRU/DP versus the respective rows of the PRU and of the DU, for the lowest values. Such a comparison shows that firstly the PRU and then the DU are more likely, for any of the forecasting models to have

a stock-out. This output is at odds with what one would expect from a part like the LRU/DP which was specified in the simulation model to have the lowest reliability among the parts (Section 6.3.1). This is a very important observation that has come as an indirect output of the present research, and is discussed in more detail in Appendix B.

7.3.5 Discussion

Again, even though MASE's outputs were clear and showed that the unsupervised BN with only the component of interest each time to forecast for (BN 1) had the best performance in all comparisons, the conclusion was not so clear with the accuracy implication metrics.

Regarding the latter, the decision maker receives more information to use, including how much more the Operational Availability increases for more Holding Volume, as well as which component part is more prone to create a stock-out problem and thus is more likely to be responsible for the lack of Operational Availability.

The reasons for this discrepancy between the accuracy metric outputs and the accuracy implication metrics are concerned with the bias of the forecasting model, the actual data, the appropriateness of the used demand distribution model and the spread parameter. The latter two are choices that were made following the approaches that are more usually taken in practice: a normal distribution model and the root mean square error (Strijbosch, Syntetos, Boylan, and Janssen, 2011; R. H. Teunter and Duncan, 2009). As was also demonstrated in Figure 7-6, Figure 7-7 and Figure 7-8 and is discussed further in Appendix B, the spread of the results also depends on the demand context and thus – given the amount of influence it has on the outputs – it also needs to be modelled as was done in the present research with the location parameter.

7.4 Conclusions

Chapter 7 presented the results of the two FPP scenarios that were simulated and studied. In both scenarios the Mean Absolute Scaled Error ($MASE$) accuracy metric suggested that the unsupervised approach to developing a BN DAG

performed better than the other BN DAG development approaches, as well as the logistic regression and the SME adjusted Single Exponential Smoothing (SES) forecast.

However, the results were not as conclusive when the accuracy implication metrics were studied, mainly because apart from the accuracy of the forecast, the variance is also influential.

Finally, regarding the FPP itself, a number of observations were made that are further studied in the accompanying Appendix B.

8 CONCLUSIONS AND FUTURE RESEARCH

8.1 Introduction

This chapter starts by reviewing the thesis aim and objectives, and also briefly presents the way that these were addressed. The chapter then proceeds in reviewing the results and the contributions made for the studied FPP. Furthermore, the limitations of the study are also reviewed. These limitations, along with a number of additional points are discussed and presented as areas for future work.

8.2 Review of the Thesis' Aim and Objectives

The aim of this research has been to study the demand context which exists during the final phase of support operations and moreover investigate the benefits of using Bayesian Networks as spares' demand forecasting models within that context. The problem examined was that of a support provider, when the customer announces that the operated systems will no longer be supported. This general problem was named the Final Phase Problem (FPP). In Chapter 2, the distinctive nature of this problem was established and its relationship to similar problems (newsvendor and last time buy problems) discussed. More specifically, the problem was examined in the context of military operations that are about to start their withdrawal phase. Bayesian Network models were chosen to model demand in this type of problem for the following reasons (Section 1.3):

- As shown in Section 2.4, there has been growing interest in and successful application of BNs in areas like reliability, dependability and maintenance, which are related to the problem of demand forecasting within the FPP context
- The changes in the support and operational context of the final phase are a cause of high uncertainty to the support providers, and the BNs have powerful features that facilitate modelling uncertainty
- The BNs' graphical component maps the associative relationships among the variables of the study, which as shown in Sections 7.2.3 and 7.3.3 and

further examined in Appendix B can help identify relationships among the variables that are not easily captured by intuition only

- The BNs' structure as a joint probability distribution provides the additional ability to answer questions (queries) about the relationships among any other subsets of the participating variables

Four different methods of BN structure development were employed and their forecasts compared:

- Unsupervised machine learning
- Elicitation of the BN structure from experts' knowledge
- Hybrid development of the BN structure using the expert knowledge as a prior structure and adding a machine learning algorithm that builds upon the elicited structure
- Hybrid development of the BN structure using the expert knowledge and adding a machine learning algorithm that uses that structure as a starting model

To benchmark results, the following commonly employed forecasting models were also developed:

- A logistic regression for the modelling of the probability of component's failure
- A Single Exponential Smoothing (SES) algorithm that provides predictions to decision makers based on past demand in order for them to adjust given their knowledge of changing demand context factors

In order to make meaningful comparisons, accuracy and accuracy implication measures were reviewed and their suitability was assessed for the FPP type of problems.

The data that were used for the development of the forecast models, but also for their evaluation, were acquired from a number of computer simulated scenarios (Sections 6.2, 6.3, 6.4, 7.2 and 7.3). The scenarios that were examined, included the systems that had components with different reliabilities and different repair practices. The research interest was in the forecast of the experienced failure rate

of these components and therefore these were the main variables of interest of the BN models.

8.3 Review and Contributions

Using the assumption of close partnerships among the organisations that operate the Support Chain (SC), the thesis has assumed that the modeller has access to different areas to collect data. This assumption was justified mainly by the type of support contracts that can exist (availability type of contracts), and also by the literature's reported trend that today's organisations tend to have closer relationships and thus increased visibility (Section 1.2). Consequently, this research intended to exploit the type of data records likely to be found in the logbooks of the different nodes of the SC including the Operations, and to use them in order to develop the forecast models.

The use of the data from the nodes in the SC system is not only a convenience but, in the case of major changes to the support requirements and resources, it is also a necessity. As was shown both from a number of the research papers in the literature review and subsequently reinforced from SME's interviews, a dominant issue in the SC is the lack of a cross-node view and therefore a lack of an overall understanding of the SC (Section 3.2). On the other hand, the data are inevitably the results of the synergies of the SC activities and thus, they do include an objective pool of SC-wide information.

The specific problem examined in the thesis, the Final Phase Problem (FPP), has not been examined in the literature as was shown in Chapter 2 where its relationships with Newsvendor problems and Last Time Buy problems were discussed. The literature review also identified a set of factors influencing spares demand. In Chapter 3, some additional influential factors were identified from interviews with subject matter experts. Additionally, a number of conceptual models were developed that can help in the identification of the specific factors to use in for spares' demand forecast models (Section 3.3). The use of one of these conceptual models was demonstrated in Chapter 6 where the development of the simulation model was discussed.

Apart from the lack of a cross-nodes' view, an additional challenge identified has to do with what can be learnt from the past data. Both scenario simulated experiments assumed that the particular demand pattern and support chain configuration experienced in the final phase did not exactly repeat one found in an earlier phase of operations (end of Section 7.2.3.6 and Appendix B). This demonstrates one of the benefits of using models like BNs which can take account of the factors that are responsible for the changes in the demand context (Section B.1 of Appendix B).

The above highlight two additional novel outputs of the thesis. Firstly, the changes in the support and operational context that can take place during the FPPs can be such that the use of past demand patterns to provide forecasts can be misleading. Secondly, models like BNs can identify those factors that are most influential and thus should be included as variables in the forecast of the demand.

Regarding the thesis objectives related to the BNs' structure development (Section 1.4), one of the early identified outputs was that the chosen type of BNs' unsupervised structure learning algorithm should be a score-based and not a constraint-based one. The research verified this observation by tests using data from the scenario, and the thesis presented two reasons for preferring the score-based approach (Section 4.3.2.3). Firstly, in the datasets of interest to the present research, a number of variables had values that were comparatively very rare, and this created a serious problem of increased false negative outputs in the tests of independence (the null hypothesis) which are used in the constraint-based algorithms. Secondly, the MBDeu score metric used in the score-based algorithms has been developed using a Bayesian parametrisation with an assumption of prior values for the parameters and then averaging over the resulting members of a family of Dirichlet distributions. This approach works as a safeguard against parameter overfitting.

All BN structures had their NPTs calculated using the dataset acquired from the simulated scenarios, via the Bayesian estimation approach as compared to the maximum likelihood estimation. The reason is again that certain variable

combinations of the joint probability distributions were empty of counts in the dataset due to a number of the variables having very rare values.

Two simulation experiments were used to compare the models: one with a simple type of supported system and another with a more elaborate one. The first scenario compared four BN models, a logistic regression and an expert-elicited forecasting approach, while the second scenario had eight BN models and a logistic regression. The first scenario's out of sample data were 144 different alternative cases, each replicated 100 times (14,400 outputs), while the second were 512 different alternative cases, each replicated again 100 times (51,200 outputs). The research did not proceed further with the expert-elicited forecasting approach because in the first scenario where it was applied, it gave comparatively poor outputs.

The study on the accuracy metrics and on the accuracy implication metrics that was performed in order to choose from for the evaluation of the forecast models, provided two additional outputs that the author considers to be novel.

Firstly, the algebraic method of analysis of the accuracy metrics, apart from confirming literature's empirically identified problems of certain accuracy measures, it also revealed some more, while it additionally demonstrated that not considering the algebraic analysis can lead to badly defended suggestions (Section 5.2.2). Particularly, the analysis showed that the Absolute Percentage Error (APE) can be problematic when there are datasets that occasionally have very small or very large values as compared to a forecast method's resulting errors (Section 5.2.2.3). The analysis also showed that for the symmetric Absolute Percentage Error ($sAPE$) (Sections 5.2.2.4, 5.2.2.5 and 5.2.2.6), apart from the already documented lack of symmetry, it is also unable to distinguish the performance among competing forecast models for a large range of error values. Moreover, the analysis showed that its three variants are identical in the errors' range from -∞ and up to an error value equal to the data point, while neither of the three should be used for errors larger than that.

The second related novel contribution of this thesis was about the accuracy implication metrics that are required in order to evaluate the forecast models in

the FPPs (Sections 5.2.5, 5.2.6). It was shown that for the cases that were examined, the commonly applied average spares Holding volume and average spares Backlogged volume need to be complimented by the end of period corresponding values and the probability of stock-out during the final phase period. Additionally, it was suggested that the measures of the support system's effectiveness approximated by the average and the end-of period volume of components' backlogs can be biased when averaging over datasets of different parts which can be strongly dependent. Given the assumption of having access to the level of operations for the acquisition of data, it was recommended that this potential problem can be mitigated by replacing the spares Backlogged volume as a measure of effectiveness by the Operational Availability of the supported systems.

In both the scenarios that were examined, the $MASE$ accuracy measure that was eventually adopted suggested the unsupervised machine learning approach to BN structure development as best. However, the evaluation output was not as clear when the accuracy implication measures were used. Even though that BN did well in these measures too, in the first scenario the logistic regression and in the second scenario the elicited BN produced better Operational Availability but for higher inventory investment both on average and at the end of the final phase.

The research showed that the reason for this discrepancy between the accuracy and the accuracy implication measures' results is that the first compare only the models' forecasts of the demand's location parameter, while the second incorporate the effects of the spread and the distribution model (Sections 7.2.5, 7.3.5).

From the modelling perspective, a number of benefits were identified from the development and use of BNs in the FPP cases. One such benefit has to do with the predictor variables' values. A common way for the models to provide forecasts is by using the known, fixed values of their explanatory variables (Gelman et al., 2014, p.5) and then producing the forecast value of the response variable. However, there can be predictors of interest, such as the workload level at the repair shop ($WWkld$) (Table 6-1) whose values are not certain/fixed at the time

that the forecast is required. This is where models like BNs provide an important benefit. Even though other modelling approaches can be modified to accommodate such a modelling challenge (see for example the logistic regression models of Sections A.1.3 and A.3.3 in Appendix A) this situation is quite common in models like BNs. They can still use variables whose values are not known by marginalising them out.

The research also revealed that there are dependencies among the components that can affect a number of the support factors. The simulated scenarios and subsequent experiments showed that the experienced times between failures and the logistic delay times of all the components are affected by specific influential ones. This is where the BN DAGs can provide useful information, since they can highlight what these influential parts might be (Section 7.3.4 and Appendix B). In addition, the most influential factors that were identified were (see also DAGs in Appendix A):

1. The factors related to the repair workload ($BWkld$, $WWkld$)
2. The environmental conditions (Env)
3. The operational demand ($OpDem$)
4. The number of units (xNU)
5. The number of mechanics (xNM)

Also studied were the differences that could be identified from the simulation of the two scenarios and what intuition could be derived from the models, the results of which can be also considered novel. It was realised that the transition from a simple system to a more elaborate one had an effect on the spread of the demand distribution, on the shape and on the negativity of its skewness. Furthermore, the latter fact made a component with non-intermittent demand to become intermittent (Appendix B).

This observation can have multiple implications if it is not correctly dealt with. Firstly, the non-zero likelihood of having intermittent future demand can mislead models that rely on only simple time-series data to predict for the final phase of operations, if just by chance – as happened in the scenarios of Section 7.3 – the in-sample dataset is not intermittent. Secondly, it was also shown (Appendix B)

that either the estimation or the evaluation of the Operational Availability of subsystems could be miscalculated if those components that are mostly influential for the waiting times are not considered within the terms of the Operational Availability function. Moreover, it is suggested that the latter risk is higher when a system undergoes modifications or upgrades on some of its components, since in such cases all components' calendar times to failure could be affected.

What was shown is that as compared at least to the regression models, using the BNs, either as an overall model with all the parts or using an individual model for each part, produces DAGs that can highlight in a very efficient way which factors are more influential and what other factors they influence. This output can then be used to inform priorities for inventory management decisions.

The following list summarises the thesis' outputs related to the aims and objectives (Section 1.4):

1. Regarding the BNs' structure development:
    a. For the development of the unsupervised BNs, the score-based algorithm was preferred to the constrained-based ones
    b. All the NPTs were calculated from data using a Bayesian estimation approach which was preferred to the maximum likelihood estimation
    c. The BNs that were built using the unsupervised learning algorithm performed better in the accuracy metric comparison. However, when the comparison used the accuracy implication measures, the results were not as conclusive, mainly because the higher effectiveness accuracy implication output was given at the expense of lower efficiency
    d. It should be expected that at the beginning of the final phase period, the values of a number of influential variables will not be known with certainty. Even though other modelling approaches can potentially be modified to use those variables' probability distributions instead, such a situation is intrinsic in the calculations of the BNs, and thus

they can more efficiently handle the problem and thus provide with a forecast

e. The BNs' directed acyclic graphs, particularly the one associated with the unsupervised learning algorithm, can identify those components whose failure rate is more influential for the support system. Having such an output can help in a number of decisions, ranging from the development of other forecast models to managing the inventory. This can also offer practitioners useful information about which variables are key.

2. Regarding the study of the demand context during the FPP:
    a. The use of data from different nodes of the Support Chain can reveal cross-nodes' synergies that can affect the demand for spares, and which are not readily perceived by SMEs and practitioners. These effects can be captured by models like BNs
    b. The identification of the possible factors that can be influential on the modelling of the demand for spares can be facilitated by the use of one of the conceptual models described in Section 3.3
    c. Due to the difference in the components' anticipated failure rates, accuracy metrics like $MASE$ (Section 5.2.3) should be preferred to those based on functions like the AE, SE, APE or $sAPE$
    d. The simulation experiments showed that the patterns of the demand for spares that existed prior to the final phase are likely not to be repeated due to the changes in the support system, and thus care should be taken not to mislead decision making
    e. The simulation experiments also showed that an increase in the complexity of the Equipment Breakdown Structure of a system can change the components' demand behaviour. Again models like BNs can help in the identification of which components are more influential for such changes

Finally, the general operationalised suggestions that would be provided to an operations manager overseeing the FPP in a critical mission, are the following:

1. It would be advised to maintain and oversee data on the factors presented in Table 6-1, or at least on the influential ones presented earlier in the current Section 8.3
2. The manager should be aware that the relationship between the numbers of resources (number of units, number of mechanics) is not guaranteed to be linear in the number of failures. Consequently, ways that can capture these non-linearities should be developed to facilitate informed decisions
3. On the same topic, developing models like BNs that can capture and present the associations among the variables can highlight dependencies and drivers that are not always apparent

8.4 Limitations

An apparent limitation of the research was that the data that were used came from simulated scenarios. However, as also discussed in Section 6.2 the approach followed had a number of benefits, including the ability to study the situation and also test the models by the use of simulation experiments. Furthermore, a simulation approach helps to identify what type of data from the support chain is most important to have. This can help to influence future practice in the collection and sharing of such data.

Another limitation of the study concerns the assumptions used in the simulated scenarios. These assumptions included the inventory policy of (S, S-1) which is a policy commonly applied when supporting high-value systems like the ones operated by the Armed Forces (Sherbrooke, 2004, sec.1.2), perfect fault detections and perfect repairs.

While the supported system in the second scenario did include a number of different components, each with different reliability and different repair process, real systems are even more complex.

Furthermore, the operational demand that was assumed in the scenarios had only two levels, while in a real situation the number of different mission configurations can be higher.

A limitation to the BN models that were developed was the need to discretise certain variables. This is a limitation, especially when at the time of the forecast there is information about a variable's value that is not included in the set of its multinomial discretised mapping. On the other hand, in the scenarios that were explored, whenever such a situation emerged, the respective variable was marginalised out, while there are also a number of benefits from discretisation (Section 4.4).

8.5 Future Work

Regarding ways in which this research can move forward, these can be seen under the following four categories:

- The support
- The supported system
- Its usage
- The forecast models

The support scenarios should have some of their assumptions changed. These include the assumption of the support policies, including the lack of preventive maintenance and the (S, S-1) resupply and repair policies, while the effects of imperfect fault-detections and repairs on the demand models should also be explored.

Furthermore, regarding the operated and supported systems' complexity, they should include more components, with even larger differences in their inherent reliabilities and also ageing effects (see also discussion about a system with more elaborate Equipment Breakdown Structure (EBS) at the end of Section B.1).

Additionally, the Operational context should be more complex than the assumed two states. Nevertheless, the conclusions resulting from the methods and the observations should be test-validated using real-life datasets.

Moreover, regarding the forecast modelling itself, it would be interesting to see whether a model of not only the demand location parameter, but also of the spread or even the skewness of the demand distribution would result in more accurate demand forecasts, especially when accuracy implication metrics are

used. On this topic, it would be interesting to study other modelling approaches that do not require the data discretisation pre-processing step of the current thesis' BNs. More specifically, relevant approaches of interest are BNs with dynamic discretisation, hierarchical Bayesian regression models or more advanced generalised regression models like the Generalised Additive Models for Location, Scale and Shape (GALMSS). Furthermore, as was suggested by the simulation outputs (Sections 7.3.5), future research could benefit if alternative demand models to the commonly applied normal distribution were also explored.

The Final Phase Problem that this thesis studied is an important, challenging problem that has been overlooked by the literature. The research examined its features and studied the development of BNs for use as demand forecasting models. This study of the BN models and also the route taken of using simulated scenarios provided a number of novel outputs that can shed light on support decisions when faced with this type of problem.

REFERENCES

Abdel-Malek, L.L. and Montanari, R. (2005) ‘An analysis of the multi-product newsboy problem with a budget constraint’, *International Journal of Production Economics*, 97, pp. 296–307.

Alwan, L.C., Xu, M., Yao, D.Q. and Yue, X. (2016) ‘The Dynamic Newsvendor Model with Correlated Demand’, *Decision Sciences*, 47(1), pp. 11–30.

Andrawis, R.R. and Atiya, A.F. (2009) ‘A New Bayesian Formulation for Holt ’s Exponential Smoothing’, *Journal of Forecasting*, 28(October), pp. 218–234.

Armstrong, J.S. and Collopy, F. (1992) ‘Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons By J. Scott Armstrong and Fred Collopy Reprinted with permission form’, *International Journal of Forecasting*, 8(1), pp. 69–80.

Armstrong, J.S. and Fildes, R. (1995) ‘Correspondence on the selection of error measures for comparisons among forecasting methods’, *Journal of Forecasting*, 14(1), pp. 67–71.

Aven, T. (2016) ‘Ignoring scenarios in risk assessments: Understanding the issue and improving current practice’, *Reliability Engineering and System Safety*, 145 Elsevier, pp. 215–220.

Axsater, S. (2006) *Inventory Control*.Second Ed.Frederick S. Hillier (Stanford University) (ed.) Lund, Swenden: Springer.

Banks, J., Carson, J.S.I., Nelson, B.L. and Nicol, D.M. (2001) *Discrete-Event System Simulaton*. Prentice Hall.

BBC (2016) *Timeline: Iraq War*. Available at: www.bbc.co.uk/news/magazine-36702957 (Accessed: 8 February 2018).

BBC (2018) *Afghanistan profile - Timeline*. Available at: www.bbc.co.uk/news/world-south-asia-12024253 (Accessed: 8 February 2018).

BBC (2010) *BAE Woodford Nimrod plant to ‘close early’*. Available at:

www.bbc.co.uk/news/uk-england-manchester-11945811 (Accessed: 8 February 2018).

BBC (2011) *Scrapping RAF Nimrods 'perverse' say Military Chiefs*., *BBC News* Available at: www.bbc.co.uk/news/uk-england-12294766 (Accessed: 23 June 2018).

Behfard, S., Van Der Heijden, M.C., Al Hanbali, A. and Zijm, W.H.M. (2015) 'Last time buy and repair decisions for spare parts', *European Journal of Operational Research*, 244(2), pp. 498–510.

Berk, E., Gürler, Ü. and Levine, R. a. (2007) 'Bayesian demand updating in the lost sales newsvendor problem: A two-moment approximation', *European Journal of Operational Research*, 182, pp. 256–281.

Boulkaibet, I., Belarbi, K., Bououden, S., Marwala, T. and Chadli, M. (2017) 'A new T-S fuzzy model predictive control for nonlinear processes', *Expert Systems with Applications*, 88 Elsevier Ltd, pp. 132–151.

Boutselis, P. and McNaught, K. (2018) 'Using Bayesian Networks to Forecast Spares Demand from Equipment Failures in a Changing Service Logistics Context', *International Journal of Production Economics*

Boylan, J.E. and Syntetos, A.A. (2006) 'Accuracy and Accuracy Implication Metrics for Intermittent Demand', *Foresight: International Journal of Applied Forecasting*, (4), pp. 39–42.

Budnitz, R.J., Apostolakis, G., Boore, D.M., Cluff, L.S., Coppersmith, K.J., Cornell, C.A. and Morris, P.A. (1998) 'Use of technical expert panels: Applications to Probabilistic Seismic Hazard Analysis', *Risk Analysis*, 18(4), pp. 463–469.

Cain, J. (2001) *Planning Improvements in Natural Resources Management*. Centre for Ecology and Hydrology, Wallingford.

Cameron, G. (2010) *TRIZICS*. CreateSpace.

Carrizosa, E., Olivares-Nadal, A. V. and Ramirez-Cobo, P. (2016) 'Robust newsvendor problem with autoregressive demand', *Computers and Operations*

Research, 68 Elsevier, pp. 123–133.

Casimir, R.J. (2002) 'The Value of Information in the Newsvendor Problem', *Omega*, 30, pp. 45–50.

Cavdar, S.C. and Aydin, A.D. (2015) 'An Empirical Analysis for the Prediction of a Financial Crisis in Turkey through the Use of Forecast Error Measures', *Journal of Risk and Financial Management*, 8(3), pp. 337–354.

Chandrasekhar, A., Larreguy, H. and Xandri, J.P. (2015) 'Testing Models of Social Learning on Networks: Evidence from a Lab Experiment in the Field', *Physiological Research*, 64(6), pp. 897–905.

Chen, F.Y., Yan, H. and Yao, L. (2004) 'A newsvendor pricing game', *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, 34(4), pp. 450–456.

Chen, L.H. and Chen, Y.C. (2010) 'A multiple-item budget-constraint newsboy problem with a reservation policy', *Omega*, 38(6) Elsevier, pp. 431–439.

Chen, Z. and Yang, Y. (2004) *Assessing Forecast Accuracy Measures* Iowa State University,

Chmielewski, M.R. and Grzymala-Busse, J.W. (1996) 'Global discretization of continuous attributes as preprocessing for machine learning', *International Journal of Approximate Reasoning*, 15(4), pp. 319–331.

Choi, T.M., Li, D. and Yan, H. (2004) 'Optimal single ordering policy with multiple delivery modes and Bayesian information updates', *Computers and Operations Research*, 31(12), pp. 1965–1984.

Christopher, M. (2016) *Logistics & Supply Chain Management*. FT Publishing International.

Christopher, M. and Lee, H. (2004) 'Mitigating supply chain risk through improved confidence', *International Journal of Physical Distribution & Logistics Management*, 34(5), pp. 388–396.

Christopher, M. and Peck, H. (2004) 'Building the Resilient Supply Chain',

International Journal of Logistics Management, 15(2), pp. 1–13.

Chu, P.-Y., Chang, K.-H. and Huang, H.-F. (2011) 'The Role of Social Mechanisms in Promoting Supplier Flexibility', *Journal of Business-to-Business Marketing*, 18, pp. 155–187.

Clemen, R.T. and Winkler, R.L. (1999a) 'Combining Probability Distributiond from experts in Risk Analysis', *Risk Analysis*, 19(2), pp. 155–156.

Clemen, R.T. and Winkler, R.L. (1999b) 'Combining Probability Distributions from Experts in Risk Analysis', *Risk Analysis*, 19(2), pp. 187–203.

Cohen, M.A., Kleindorfer, P.R., Lee, H.L. and Pyke, D.F. (1992) 'Multi-Item Service Constrained (s,S) Policies for Spare Parts Logistics Systems', *Naval Research Logistics*, 39(May), pp. 561–577.

Cohen, M.A., Zheng, Y.-S. and Agrawal, V. (1997) 'Service parts logistics: a benchmark analysis', *IIE Transactions*, 29(8), pp. 627–639.

Coleman, C.D. and Swanson, D.A. (2007) 'On MAPE-R as a measure of estimation and forecast accuracy', *Journal of Economic and Social Measurement*, 30(January), pp. 219–233.

Committee on Force Multiplying Technologies for Logistics Support to Military Operations and Board on Army Science and Technology, . (2014) *Force Multiplying Technologies for Logistics Support to Military Operations*. National Academy of Sciences.

Cooper, G.F. and Herskovits, E. (1992) '*A Bayesian Method for the Induction of Probabilistic Networks from Data*', 347, pp. 309–347.

Cooper, G.F. and Yoo, C. (1999) Causal discovery from a mixture of experimental and observational data *Proceedings Fifthteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*.

Crandal, B., Klein, G. and Hoffman, R. (2006) *Working Minds - A Practitioner's Guide to Cognitive Task Analysis*. Massachusetts Institute of Technology Press.

Dash, D. and Druzdzel, M. (2002) Robust independence testing for constraint-

based learning of causal structure *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI'03)*.

Davenport, T.H. and Prusak, L. (2000) Working knowledge. How organizations manage what they know *Ubiquity*. Harvard Business School Press,

Davydenko, B.A. and Fildes, R. (2016) 'Forecast Error Measures : Critical Review and Practical Recommendations', in *Business Forecasting: Practical Problems and Solutions*. , pp. 238–250.

Defence Committee (2012) *Fifth Report: Future Maritime Surveillance - Defence Committee Contents 3 - The Capability Gap.*, *UK Parliament* Available at: publications.parliament.uk/pa/cm201213/cmselect/cmdfence/110/11002.htm (Accessed: 23 June 2018).

DeGroot, M.H. (1974) 'Reaching a Consensus', *Journal of the American Statistical Association*, 69(345), pp. 118–121.

Dekker, R., Pinçe, Ç., Zuidwijk, R. and Jalil, M.N. (2013) 'On the use of installed base information for spare parts logistics: A review of ideas and industry practice', *International Journal of Production Economics*, 143(2), pp. 536–545.

DeMarzo, P.M., Vayanos, D. and Zwiebel, J. (2003) 'Persuasion bias, social influence, and uni-dimensional opinions', *Quarterly Journal of Economics*, 118(3), pp. 909–968.

Demšar, J. (2006) 'Statistical Comparisons of Classifiers over Multiple Data Sets', *Journal of Machine Learning Research*, 7, pp. 1–30.

Ding, S. and Gao, Y. (2014) 'The (σ, S) policy for uncertain multi-product newsboy problem', *Expert Systems with Applications*, 41(8) Elsevier Ltd, pp. 3769–3776.

Doguc, O. and Ramirez-Marquez, J.E. (2009) 'A generic method for estimating system reliability using Bayesian networks', *Reliability Engineering and System Safety*, 94(2), pp. 542–550.

Dombi, J., Jónás, T. and Tóth, Z.E. (2018) 'Modeling and long-term forecasting demand in spare parts logistics businesses', *International Journal of Production*

Economics, 201, pp. 1–17.

Douven, I. (2010) 'Simulating peer disagreements', *Studies in History and Philosophy of Science Part A*, 41(2), pp. 148–157.

Druzdzel, M.J. and Gaag, L.C. Van Der (2000) 'Building probabilistic networks: "Where do the numbers come from?" guest editors' introduction', *IEEE Transactions on Knowledge and Data Engineering*, 12(4), pp. 481–486.

Eaves, a H.C. and Kingsman, B.G. (2004) 'Forecasting for the ordering and stock-holding of spare parts', *Journal of the Operational Research Society*, 55, pp. 431–437.

Eppen, G.D. and Iyer, A. V (1997) 'Improved Fashion Buying with Bayesian Updates', *Operations Research*, 45(6), pp. 805–819.

Fayyad, U. and Irani, K. (1993) 'Multi-interval Discretization of Continuous-valued Attributes for Classification Learning', *Artificial Intelligence*, 13, pp. 1022–1027.

Feeney, G.J. and Sherbrooke, C.C. (1965) *A system approach to base stockage of recoverable items* The RAND Corproration,

Fenton, N. and Neil, M. (2013) *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press.

Field, A., Miles, J. and Field, Z. (2012) *Discovering Statistics Using R*. London: SAGE Publications Ltd.

Fildes, R. (1992) 'The evaluation of extrapolative forecasting methods', *International Journal of Forecasting*, 8(1), pp. 81–98.

Fildes, R., Goodwin, P., Lawrence, M. and Nikolopoulos, K. (2009) 'Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning', *International Journal of Forecasting*, 25(1) Elsevier B.V., pp. 3–23.

Fisher, M.L., Hammond, J.H., Obermeyer, W.R. and Raman, A. (1994) 'Making Supply Meet Demand in Uncertain World', *Harvard Business Review*, 72, pp. 83–93.

FlightGlobal (2006) *Farnborough: BAE wins Nimrod MRA4 contract*. Available at: www.flightglobal.com/news/articles/farnborough-bae-wins-nimrod-mra4-contract-208012/ (Accessed: 22 February 2018).

FlightGlobal (2017) *Allegiant to retire last MD-80 next November*. Available at: www.flightglobal.com/news/articles/allegiant-to-retire-last-md-80-next-november-443788/ (Accessed: 22 February 2018).

Fortuin, L. (1980) 'The All-Time Requirement of Spare Parts for Service After Sales — Theoretical Analysis and Practical Results', *International Journal of Operations & Production Management*, 1(1), pp. 59–70.

Fortuin, L. (1981) 'Reduction of the All-Time Requirement for Spare Parts', *International Journal of Operations & Production Management*, 2(1), pp. 29–37.

Franses, P.H. and Legerstee, R. (2010) 'Do experts' adjustments on model-based SKU-level forecasts improve forecast quality?', *Journal of Forecasting*, 29(3), pp. 331–340.

Frost, A. (2018) *Knowledge Management Tools - Defining Knowledge, Information, Data*. Available at: www.knowledge-management-tools.net/knowledge-information-data.html (Accessed: 17 July 2018).

van der Gaag, L.C., Renooij, S., Witteman, C., Aleman, B.M.P. and Taal, B.G. (1999) 'How to elicit many probabilities', *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*., pp. 647–654.

van der Gaag, L.C., Renooij, S., Witteveen, C.L.M., Aleman, B.M.P. and Taal, B.G. (2002) 'Probabilities for a probabilistic network: A case-study in oesophageal cancer', *Artificial Intelligence in Medicine*, 25(2), pp. 123–148.

Gardner, E.S. (1990) 'Evaluating Forecast Performance in an Inventory Control System', *Management Science*, 36(4), pp. 490–499.

Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005) 'Statistical methods for eliciting probability distributions', *Journal of the American Statistical Association*, 100(470), pp. 680–700.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2014) *Bayesian Data Analysis*. 3rd edn. Dominici, F., Faraway, J., Tanner, M. and Zidek, J. (eds.) Chapman & Hall/CRC.

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Geurts, J.H.. and Moonen, J.M.. (1992) 'On the Robustness of "Insurance Type" Spares Provisioning Startegies', *The Journal of the Operational Research Society*, 43(1), pp. 43–51.

Gigerenzer, G., Hoffrage, U., Mellers, B.A. and McGraw, A.P. (1995) 'How to Improve Bayesian Reasoning Without Instruction: Frequency Formats', *Psychological review*, 102(4), pp. 684–704.

Gilks, W.R., Thomas, A. and Spiegelhalter, D.J. (1993) 'A Language and program for complex bayesian modelling', *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(1, Special), pp. 169–177.

Golub, B. and Jackson, M.O. (2012) 'How Homophily Affects the Speed of Learning and Best-Response Dynamics', *Quarterly Journal of Economics*, (2006), pp. 1287–1338.

Gonzalez-Abril, L., Cuberos, F.J., Velasco, F. and Ortega, J.A. (2009) 'Ameva: An autonomous discretization algorithm', *Expert Systems with Applications*, 36(3 PART 1) Elsevier Ltd, pp. 5327–5332.

Goodwin, P. and Lawton, R. (1999) 'On the asymmetry of the symmetric MAPE', *International Journal of Forecasting*, 15(3, 2), pp. 405–408.

Goodwin, P. and Wright, G. (2014) *Decision Analysis for Management Judgement*. 5th edn. Wiley.

Hadley, G.F. and Whitin, T.M. (1963) *Analysis of Inventory Systems*. USA: Prentice Hall.

Hamza, W., Lusito, L., Ligorio, F., Tomasicchio, G. and D'Alessandro, F. (2018) 'Wave Climate at Shallow Waters along the Abu Dhabi Coast', *Water*, 10(8), pp.

985–1003.

Hartemink, A.J. (2001) *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks*. Massachusetts Institute of Technology.

Heckerman, D., Geiger, D. and Chickering, D. (1995) 'Learning Bayesian networks: the combination of knowledge and statistical data', *Machine Learning*, 20, pp. 197–243.

Hegselmann, R. and Krause, U. (2002) 'Opinion Dynamics and Bounded Confidence', *Simulation*, 5(3), pp. 1–33.

Hill, R.M. (1997) 'Applying Bayesian methodology with a uniform prior to the single period inventory model', *European Journal of Operational Research*, 98, pp. 555–562.

Hong, J.S., Koo, H.Y., Lee, C.S. and Ahn, J. (2008) 'Forecasting service parts demand for a discontinued product', *IIE Transactions (Institute of Industrial Engineers)*, 40(7), pp. 640–649.

Hoover, J. (2006) 'Measuring Forecast Accuracy: Omissions in Today's Forecasting Engines and Demand-Planning Software', *Foresight: International Journal of Applied Forecasting*, (4), pp. 32–35.

Hosseini, S. and Barker, K. (2016) 'A Bayesian network model for resilience-based supplier selection', *International Journal of Production Economics*, 180 Elsevier, pp. 68–87.

Hryniewicz, O. and Kaczmarek, K. (2016) 'Bayesian analysis of time series using granular computing approach', *Applied Soft Computing Journal*, 47 Elsevier B.V., pp. 644–652.

Huan Liu and Setiono, R. (1997) 'Chi2: feature selection and discretization of numeric attributes', *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*., pp. 388–391.

Huang, D., Zhou, H. and Zhao, Q.H. (2011) 'A competitive multiple-product

newsboy problem with partial product substitution', *Omega*, 39(3) Elsevier, pp. 302–312.

Hyndman, R.J. (2006) 'Another Look at Forecasting Metrics for Intermittent Demand', *Foresight: International Journal of Applied Forecasting*, (4), pp. 43–46.

Hyndman, R.J. (2015) 'Measuring Forecast Accuracy', in *Business Forecasting: Practical Problems and Solutions*. John Wiley & Sons, pp. 177–184.

Hyndman, R.J. (2014) *Errors on Percentage Errors*., *Hyndsight blog* Available at: robjhyndman.com/hyndsight/smape/ (Accessed: 28 June 2018).

Hyndman, R.J. and Koehler, A.B. (2006) 'Another look at measures of forecast accuracy', *International Journal of Forecasting*, 22(4), pp. 679–688.

Ide, J. and Cozman, F. (2002) 'Random generation of Bayesian networks', *Advances in Artificial Intelligence*, , pp. 366–376.

Inderfurth, K. and Kleber, R. (2013) 'An advanced heuristic for multiple-option spare parts procurement after end-of-production', *Production and Operations Management*, 22(1), pp. 54–70.

Inderfurth, K. and Mukherjee, K. (2008) 'Decision support for spare parts acquisition in post product life cycle', *Central European Journal of Operations Research*, 16(1), pp. 17–42.

Jensen, F. V. and Nielsen, T.D. (2007) *Bayesian Networks and Decision Graphs*. Springer.

Jiang, H. (2013) *Key Findings on Airplane Economic Life*. Available at: www.boeing.com/assets/pdf/commercial/aircraft_economic_life_whitepaper.pdf (Accessed: 22 February 2018).

Johnston, F.R., Boylan, J.E. and Shale, E.A. (2003) 'An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items', *Journal of the Operational Research Society*, 54(8), pp. 833–837.

Jones, B., Jenkinson, I., Yang, Z. and Wang, J. (2010) 'The use of Bayesian

network modelling for maintenance planning in a manufacturing industry', *Reliability Engineering and System Safety*, 95(3) Elsevier, pp. 267–277.

Jouffe, L., Weber, P. and Munteanu, P. (2004) 'Dynamic Bayesian Networks modelling the dependability of systems with degradations and exogenous constraints', *11th IFAC Symposium on Information Control Problems in Manufacturing*

Kahneman, D., Slovic, P. and Tversky, A. (1974) 'Judgment under uncertainty: heuristics and biases', *Science*, 185(4157), pp. 1124–1131.

Kang, C.W. and Golay, M.W. (1999) 'A Bayesian belief network-based advisory system for operational availability focused diagnosis of complex nuclear power systems', *Expert Systems with Applications*, 17(1), pp. 21–32.

Kaplan, A., Skogstad, A.L. and Girshick, M.A. (1950) 'The Prediction of Social and Technological Events', *The Public Opinion Quarterly*, 14(1), pp. 93–110.

Kennedy, W.J., Patterson, W. and Fredendall, L.D. (2002) 'An overview of recent literature on spare parts inventories', *International Journal of Production Economics*, 76(2), pp. 201–215.

Kerber, R. (1992) 'ChiMerge: Discretization of numeric attributes', *Proceedings of the tenth national conference on Artificial intelligence*, , pp. 123–128.

Khouja, M. (1999) 'The single-period (news-vendor) problem: Literature review and suggestions for future research', *Omega*, 27, pp. 537–553.

Khouja, M. and Mehrez, A. (1996) 'A multi-product constrained newsboy problem with progressive multiple discounts', *Computers and Industrial Engineering*, 30, pp. 95–101.

Kim, B. and Park, S. (2008) 'Optimal pricing, EOL (end of life) warranty, and spare parts manufacturing strategy amid product transition', *European Journal of Operational Research*, 188(3), pp. 723–745.

Kim, H. (2012) *Package discretize in R* 1.0-1.

Kim, T.Y., Dekker, R. and Heij, C. (2017) 'Spare part demand forecasting for

consumer goods using installed base information', *Computers and Industrial Engineering*, 103 Elsevier Ltd, pp. 201–215.

Kirkup, J. (2010) *£3.6 billion Nimrods Dismantled for Scrap*., *The Telegraph*

Klassen, R.D. and Flores, B.E. (2001) 'Forecasting practices of Canadian firms: Survey results and comparisons', *International Journal of Production Economics*, 70(2), pp. 163–174.

Kleber, R., Schulz, T. and Voigt, G. (2012) 'Dynamic buy-back for product recovery in end-of-life spare parts procurement', *International Journal of Production Research*, 50(6), pp. 1476–1488.

Koehler, A.B. (2001) 'The Assymetry of the sAPE Measure and other Comments on the M3-competition', *International Journal of Forecasting*, 17, pp. 570–574.

Van Kooten, J.P.J. and Tan, T. (2009) 'The final order problem for repairable spare parts under condemnation', *Journal of the Operational Research Society*, 60(10), pp. 1449–1461.

Korb, K.B. and Nicholson, A.E. (2004) *Bayesian Artificial Intelligence*. London: Chapman & Hall/CRC.

Kourentzes, N. (2013) 'Intermittent demand forecasts with neural networks', *International Journal of Production Economics*, 143(1) Elsevier, pp. 198–206.

Kraiselburd, S., Narayanan, V.G. and Raman, A. (2009) 'Contracting in a Supply Chain with Stochastic Demand and Substitute Products', *Production and Operations Management*, 13(1), pp. 46–62.

Krikke, H. and Van Der Laan, E. (2011) 'Last Time Buy and control policies with phase-out returns: A case study in plant control systems', *International Journal of Production Research*, 49(17), pp. 5183–5206.

Lancaster, D.D. (2005) *Developing a Fly-Away-Kit (FLAK) to support Hastily Formed Networks (HFN) for Humanitarian Assistance and Disaster Relief*. Naval Postgraduate School (NPS).

Langseth, H. (1998) 'Analysis of survival times using Bayesian networks',

Proceedings of the 9th European Conference on Safety and Reliability. Rotterdam: A.A. Balkema, pp. 647–654.

Langseth, H., Haugen, K. and Sandtorv, H. (1998) 'Analysis of OREDA data for maintenance optimisation', *Reliability Engineering & System Safety*, 60(2), pp. 103–110.

Langseth, H. and Portinale, L. (2007) 'Bayesian networks in reliability', *Reliability Engineering and System Safety*, 92(1), pp. 92–108.

Lariviere, M. a. and Porteus, E.L. (1999) 'Stalking Information: Bayesian Inventory Management with Unobserved Lost Sales', *Management Science*, 45(3), pp. 346–363.

Laskey, K. and Mahoney, S. (1998) Network Fragments for Knowledge-Based Construction of Belief Networks *AAAI Technical Report SS-98-03*.

Lau, H. and Song, H. (2008) 'Multi-echelon repairable item inventory system with limited repair capacity under …', *International Journal of Inventory Research*, , pp. 67–92.

Law, A.M. and Kelton, D.W. (1991) *Simulation Modelling and Analysis*. McGraw-Hill.

Leifker, N.W., Jones, H.C. and Lowe, T.J. (2014) 'Determining optimal order amount for end-of-life parts acquisition with possibility of contract extension', *Engineering Economist*, 59(4), pp. 259–281.

Leifker, N.W., Jones, P.C. and Lowe, T.J. (2012) 'A continuous-time examination of end-of-life parts acquisition with limited customer information', *Engineering Economist*, 57(4), pp. 284–301.

Liu, H., Hussain, F., Tan, C.L.I.M. and Dash, M. (2002) '*Discretization An enabling technique*', , pp. 393–423.

Liu, W., Dou, Z., Wang, W., Liu, Y., Zou, H., Zhang, B. and Hou, S. (2018) 'Short-Term Load Forecasting Based on Elastic Net Improved GMDH and Difference Degree Weighting Optimization', *Applied Sciences*, 8(9), pp. 1603–1624.

Lodree, E.J., Kim, Y. and Jang, W. (2008) ‘Time and quantity dependent waiting costs in a newsvendor problem with backlogged shortages’, *Mathematical and Computer Modelling*, 47(1–2), pp. 60–71.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2013) *The BUGS Book A Practical Introduction to Bayesian Analysis*. CRC Press Taylor & Francis Group.

Luo, Z., Wang, J. and Chen, W. (2015) ‘A risk-averse newsvendor model with limited capacity and outsourcing under the CVaR criterion’, *Journal of Systems Science and Systems Engineering*, 24(1), pp. 49–67.

Madigan, D., York, J. and Allard, D. (1995) ‘Bayesian Graphical Models for Discrete Data’, *International Statistical Review*, 63(2), pp. 215–232.

Mahoney, S.M. and Laskey, K.B. (1996) ‘Network Engineering for Complex Belief Networks’, *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, , pp. 389–396.

Makridakis, S. (1993) ‘Accuracy measures : theoretical and practical concerns’, *International journal of forecasting*, 9, pp. 527–529.

Makridakis, S. and Hibon, M. (2000) ‘The M3-Competition: results, conclusions and implications’, *International Journal of Forecasting*, 16(4), pp. 451–476.

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (2008) *Forecasting Methods and Applications*. Wiley.

Marcot, B.G., Steventon, J.D., Sutherland, G.D. and McCann, R.K. (2006) ‘Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation’, *Canadian Journal of Forest Research*, 36(12), pp. 3063–3074.

Margaritis, D. (2003) *Learning Bayesian Network Model Structure from Data*. School of Computer Science, Carnegie Mellon University.

Martínez-Álvarez, F., Troncoso, A., Asencio-Cortés, G. and Riquelme, J. (2015) ‘A Survey on Data Mining Techniques Applied to Electricity-Related Time Series

Forecasting', *Energies*, 8(12), pp. 13162–13193.

MATLAB (2017) *Maths, Statistics and Optimisation toolbox: User's Guide - boxplot* Mathworks,

McNaught, K. and Chan, A. (2011) 'Bayesian networks in manufacturing', *Journal of Manufacturing Technology Management*, 22(6), pp. 734–747.

McNaught, K. and Zagorecki, A. (2009) 'Using dynamic Bayesian networks for prognostic modelling to inform maintenance decision making', *IEEM 2009 - IEEE International Conference on Industrial Engineering and Engineering Management*, (June), pp. 1155–1159.

Medina-Oliva, G., Weber, P., Simon, C. and Iung, B. (2009) 'Bayesian networks applications on dependability, risk analysis and maintenance', *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 2(PART 1) IFAC, pp. 215–220.

Meridiana (2018) *Commercial Aircraft: Line Maintenance*. Available at: www.meridianamaintenance.com/en-us/productservices/commercialaircraft/maintenance/linemaintenance.aspx (Accessed: 22 February 2018).

Mirzahosseinian, H. and Piplani, R. (2011) 'A study of repairable parts inventory system operating under performance-based contract', *European Journal of Operational Research*, 214(2) Elsevier B.V., pp. 256–261.

Monti, S. and Carenini, G. (2000) 'Dealing with the expert inconsistency in probability elicitation', *IEEE Transactions on Knowledge and Data Engineering*, 12(4), pp. 499–508.

Monti, S. and Cooper, G.F. (1998a) 'A Multivariate Discretization Method for Learning Bayesian Networks from Mixed Data', *Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, pp. 404–413.

Monti, S. and Cooper, G.F. (1998b) 'Learning Hybrid Bayesian Networks from Data', in Jordan, M. . (ed.) *Learnign in Graphical Models*. Springer, Dordrecht, pp. 521–540.

Moore, J.R. (1971) ‘Forecasting and Scheduling for Past-Model Replacement Parts’, *Management Science*, 18(4-Part-I), pp. 200–213.

Mrad, A. Ben, Delcroix, V., Piechowiak, S., Leicester, P. and Abid, M. (2015) ‘An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence: Uncertain evidence in Bayesian networks’, *Applied Intelligence*, 43(4), pp. 802–824.

Nagarajan, R., Scutari, M. and Lebre, S. (2013) *Bayesian Networks in R with Applications in Systems Biology*. Springer.

NATO (2017) *NATO and Afghanistan*. Available at: www.nato.int/cps/en/natohq/topics_8189.htm (Accessed: 8 February 2018).

Neapolitan, R.E. (2004) *Learning Bayesian Networks*. Prentice Hall.

Neil, M., Fenton, N. and Forey, S. (2001) ‘Using Bayesian Belief Networks to Predict the Reliability of Military Vehicles’, *Computing and Control Engineering Journal*, (February), pp. 11–20.

Neil, M., Fenton, N. and Nielson, L. (2000) ‘Building large-scale Bayesian networks’, *The Knowledge Engineering Review*, 15(3), pp. 257–284.

Neil, M., Tailor, M. and Marquez, D. (2007) ‘Inference in Hybrid Bayesian Networks using Dynamic Discretisation’, *Statistics and Computing*, 17(3), pp. 219–233.

Norwegian Petroleum Safety Authority (1997) *OREDA*. Available at: www.oreda.com/database/ (Accessed: 2 March 2018).

Nowicki, D., Kumar, U.D., Steudel, H.J. and Verma, D. (2008) ‘Spares provisioning under performance-based logistics contract: Profit-centric approach’, *Journal of the Operational Research Society*, 59(3), pp. 342–352.

Nowicki, D.R., Randall, W.S. and Ramirez-Marquez, J.E. (2012) ‘Improving the computational efficiency of metric-based spares algorithms’, *European Journal of Operational Research*, 219(2) Elsevier B.V., pp. 324–334.

O’Hagan, A. (1998) ‘*Eliciting prior beliefs in substantial practical applications*’,

47(1), pp. 21–25.

Oxford University online dictionary (2018a) *English Oxford Dictionaries*. Available at: en.oxforddictionaries.com/definition/learning (Accessed: 26 June 2018).

Oxford University online dictionary (2018b) *English Oxford Living Dictionaries*. Available at: en.oxforddictionaries.com/definition/entropy (Accessed: 12 March 2018).

Özer, Ö., Uncu, O. and Wei, W. (2007) 'Selling to the "Newsvendor" with a forecast update: Analysis of a dual purchase contract', *European Journal of Operational Research*, 182(3), pp. 1150–1176.

Page, E.H. (1994) *Simulation Modeling Methodology : Principles and Etiology of Decision Support*. Virginia Polytechnic Imstitute and State University.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 2nd edn. Branchman, R. (ed.) San Francisco: Morgan Kaufmann Publishers.

Petropoulos, F. and Kourentzes, N. (2015) 'Forecast combinations for intermittent demand', *Journal of the Operational Research Society*, 66(6), pp. 914–924.

Petropoulos, F., Makridakis, S., Assimakopoulos, V. and Nikolopoulos, K. (2014) '"Horses for Courses" in demand forecasting', *European Journal of Operational Research*, 237(1), pp. 152–163.

Pill, J. (1971) 'The Delphi Method: Substance , Context , a Critique and an Annotated Bibliography', *Socio-Economic Planning Sciences*, 5, pp. 57–71.

Pince, C. and Dekker, R. (2011) 'An inventory model for slow moving items subject to obsolescence', *European Journal of Operational Research*, 213(1), pp. 83–95.

Polatoğlu, L.H. (1991) 'Optimal order quantity and pricing decisions in single-period inventory systems', *International Journal of Production Economics*, 23, pp. 175–185.

Pourakbar, M., van Der Laan, E. and Dekker, R. (2014) 'End-of-life inventory

problem with phaseout returns', *Production and Operations Management*, 23(9), pp. 1561–1576.

Pryor, G. a. (2008) 'Methodology for estimation of operational availability as applied to military systems', *ITEA Journal*, 29(4), pp. 420–428.

Qin, Y., Wang, R., Vakharia, A.J., Chen, Y. and Seref, M.M.H. (2011) 'The newsvendor problem: Review and directions for future research', *European Journal of Operational Research*, 213(2) Elsevier B.V., pp. 361–374.

Ramoni, M. and Sebastiáni, P. (1997) *Learning Bayesian networks from incomplete databases* The Open University, Milton Keynes

Rekik, Y., Glock, C.H. and Syntetos, A.A. (2017) 'Enriching demand forecasts with managerial information to improve inventory replenishment decisions: Exploiting judgment and fostering learning', *European Journal of Operational Research*, 261(1) Elsevier B.V., pp. 182–194.

Ritchie, E. and Wilcox, P. (1977) 'Renewal theory forecasting for stock control', *European Journal of Operational Research*, 1(2), pp. 90–93.

Ross Quinlan, J. and Rivest, R.L. (1989) 'Inferring decision trees using the minimum description lenght principle', *Information and Computation*, 80(3), pp. 227–248.

Saaty, T.L. (1980) *The Analytic Hierarchy Process*.McGraw-Hill (ed.).

Sani, B. and Kingsman, B.G. (1997) 'Selecting the Best Periodic Inventory Control and Demand Forecasting Methods for Low Demand Items', *The Journal of the Operational Research Society*, 48(7), pp. 700–713.

Schneider, H. (1981) 'Effect of service-levels on order-points or order-levels in inventory models', *International Journal of Production Research*, 19(6), pp. 615–631.

Scutari, M. and Denis, J.-B. (2015a) 'Causal Bayesian Networks', in *Bayesian Networks with Examples in R*. CRC Press Taylor & Francis Group, pp. 119–122.

Scutari, M. and Denis, J.-B. (2015b) *Bayesian Networks with Examples in R*.

London: Taylor & Francis.

Sherbrooke, C.C. (2000) *Using Sorties vs . Flying Hours to Predict Aircraft Spares Demand*. Virginia US.

Sherbrooke, C.C. (2004) *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*.Hillier, F. S. (ed.) Kluwer Academic Publishers.

Sherbrooke, C.C. (1967) METRIC : A Multi-Echelon Technique for Recoverable Item Control *INFORMS: Institute for Operations Research*. CA: RAND Corporation,

Shi, J. (1997) 'A Conceptual Activity Cycled-Based Simulation Modelling Method', Andradottir, S., Healy, K. J., Withers, D. H. and Nelson, B. L. (eds.)*Proceedings of the 1997 Winter Simulation Conference*., pp. 1127–1133.

Sigurdsson, J.H., Walls, L.A. and Quigley, J.L. (2001) 'Bayesian belief nets for managing expert judgement and modelling reliability', *Quality and Reliability Engineering International*, 17(3), pp. 181–190.

Spider Financial (2018) *NumXL Spider Financial*. Available at: www.spiderfinancial.com/support/documentation/numxl/reference-manual/forecasting-performance/mdape (Accessed: 1 June 2018).

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search*. 2nd edn. Boston: MIT press.

Štěpnička, M., Cortez, P., Donate, J.P. and Štěpničková, L. (2013) 'Forecasting seasonal time series with computational intelligence: On recent methods and the potential of their combinations', *Expert Systems with Applications*, 40(6), pp. 1981–1992.

Sterman, J.D. (2000) *Business Dynamics - Systems Thinking and Modeling for a Complex World*. Boston: McGraw-Hill.

Strijbosch, L.W.G., Syntetos, A.A., Boylan, J.E. and Janssen, E. (2011) 'On the interaction between forecasting and stock control: The case of non-stationary demand', *International Journal of Production Economics*, 133(1) Elsevier, pp.

470–480.

Sujjaviriyasup, T. (2017) ‘A new class of MODWT-SVM_DE hybrid model emphasizing on simplification structure in data pre-processing: A case study on annual electricity consumptions’, *Applied Soft Computing*, 54, pp. 150–163.

Swanson, D.A., Tayman, J. and Barr, C.F. (2000) ‘A note on the measurement of accuracy for subnational demographic estimates’, *Demography*, 37(2), pp. 193–201.

Syntetos, A.A. and Boylan, J.E. (2005) ‘The Accuracy of Intermittent Demand Estimates’, *International Journal of Forecasting*, 21, pp. 303–314.

Syntetos, A.A., Boylan, J.E. and Croston, J.D. (2005) ‘On the categorization of demand patterns.’, *Journal of the Operational Research Society*, 56(5), pp. 495–503.

Syntetos, A.A., Nikolopoulos, K. and Boylan, J.E. (2010) ‘Judging the judges through accuracy-implication metrics: The case of inventory forecasting’, *International Journal of Forecasting*, 26(1) Elsevier B.V., pp. 134–143.

Syntetos, A.A., Nikolopoulos, K., Boylan, J.E., Fildes, R. and Goodwin, P. (2009) ‘The effects of integrating management judgement into intermittent demand forecasts’, *International Journal of Production Economics*, 118(1), pp. 72–81.

Systecon (2015) OPUS10 [Lecture notes] *Systecon OPUS suite*. Systecon UK,

Teunter, R.H. and Duncan, L. (2009) ‘Forecasting intermittent demand: A comparative study’, *Journal of the Operational Research Society*, 60(3), pp. 321–329.

Teunter, R.H. and Fortuin, L. (1999) ‘End-of-life service’, *International Journal of Production Economics*, 59(1), pp. 487–497.

Teunter, R.H. and Fortuin, L. (1998) ‘End-of-life service : A case study’, *International Journal of Operational Research*, 107, pp. 19–34.

Teunter, R.H. and Haneveld, K. (1998) ‘The “‘final order’” problem’, *European Journal of Operational Research*, 107, pp. 35–44.

Teunter, R.H. and Haneveld, K. (2002) ‘Inventory control of service parts in the final phase’, *European Journal of Operational Research*, 137(3), pp. 497–511.

The Seattle Times (2013) *Lockheed Martin cutting 4,000 jobs, closing plants*. Available at: https://www.cbsnews.com/news/lockheed-martin-cutting-4000-jobs-closing-plants/ (Accessed: 8 February 2018).

Thompson, P.A. (1990) ‘An MSE statistic for comparing forecast accuracy across series’, *International Journal of Forecasting*, 6(2), pp. 219–227.

US DoD (2017a) *Joint Publication 5-0: Joint Planning*. Available at: www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp5_0_20171606.pdf (Accessed: 22 June 2018).

US DoD (2017b) *Joint Publication 3-0: Joint Operations*., *Joint Publications* Available at: 10.4135/9781412952446 (Accessed: 22 June 2018).

Vairaktarakis, G.L. (2000) ‘Robust multi-item newsboy models with a budget constraint’, *International Journal of Production Economics*, 66, pp. 213–226.

Valle Dos Santos, R.D.O. and Vellasco, M.M.B.R. (2015) ‘Neural Expert Weighting: A NEW framework for dynamic forecast combination’, *Expert Systems with Applications*, 42(22) Elsevier Ltd., pp. 8625–8636.

Visual Paradigm online (2019) *Visual paradigm*., *Activity Diagram* Available at: online.visual-paradigm.com (Accessed: 24 March 2019).

Wallström, P. and Segerstedt, A. (2010) ‘Evaluation of forecasting error measurements and techniques for intermittent demand’, *International Journal of Production Economics*, 128(2), pp. 625–636.

Waters, D.C. (2011) *Quantitative Methods for Business*. 5th edn. Essex UK: Pearson Education.

Weber, P. and Jouffe, L. (2006) ‘Complex system reliability modelling with Dynamic Object Oriented Bayesian Networks (DOOBN)’, *Reliability Engineering and System Safety*, 91(2), pp. 149–162.

Weber, P., Medina-Oliva, G., Simon, C. and Iung, B. (2012) ‘Overview on

Bayesian networks applications for dependability, risk analysis and maintenance areas', *Engineering Applications of Artificial Intelligence*, 25(4), pp. 671–682.

Wiegmann, D.A. (2005) Developing a Methodology for Eliciting Subjective Probability Estimates During Expert Evaluations of Safety Interventions: Application for Bayesian Belief Networks *NASA-05-4*. University of Illinois, Illinois

Willemain, T.R. (2006) 'Forecast-Accuracy Metrics for Intermittent Demands: Look at the Entire Distribution of Demand', *Foresight: International Journal of Applied Forecasting*, (4), pp. 36–38.

Willemain, T.R., Smart, C.N. and Schwarz, H.F. (2004) 'A new approach to forecasting intermittent demand for service parts inventories', *International Journal of Forecasting*, 20(3), pp. 375–387.

Van Wingerden, E., Basten, R.J.I., Dekker, R. and Rustenburg, W.D. (2014) 'More grip on inventory control through improved forecasting: A comparative study at three companies', *International Journal of Production Economics*, 157(1), pp. 220–237.

Zamora-Martínez, F., Romeu, P., Botella-Rocamora, P. and Pardo, J. (2013) 'Towards energy efficiency: Forecasting indoor temperature via multivariate analysis', *Energies*, 6(9), pp. 4639–4659.

Zhang, R.Q., Zhang, L.K., Zhou, W.H., Saigal, R. and Wang, H.W. (2014) 'The multi-item newsvendor model with cross-selling and the solution when demand is jointly normally distributed', *European Journal of Operational Research*, 236(1) Elsevier B.V., pp. 147–159.

Zhao, Y. and Zhao, X. (2016) 'How a competing environment influences newsvendor ordering decisions', *International Journal of Production Research*, 54(1), pp. 204–214.

Zheng, M., Wu, K. and Shu, Y. (2016) 'Newsvendor problems with demand forecast updating and supply constraints', *Computers and Operations Research*, 67 Elsevier, pp. 193–206.

APPENDICES

Appendix A Forecast Models used in the Second Scenario

A.1 Models for the Forecast of the Demands in LRU

A.1.1 A Single BN with All the FRT Nodes Included

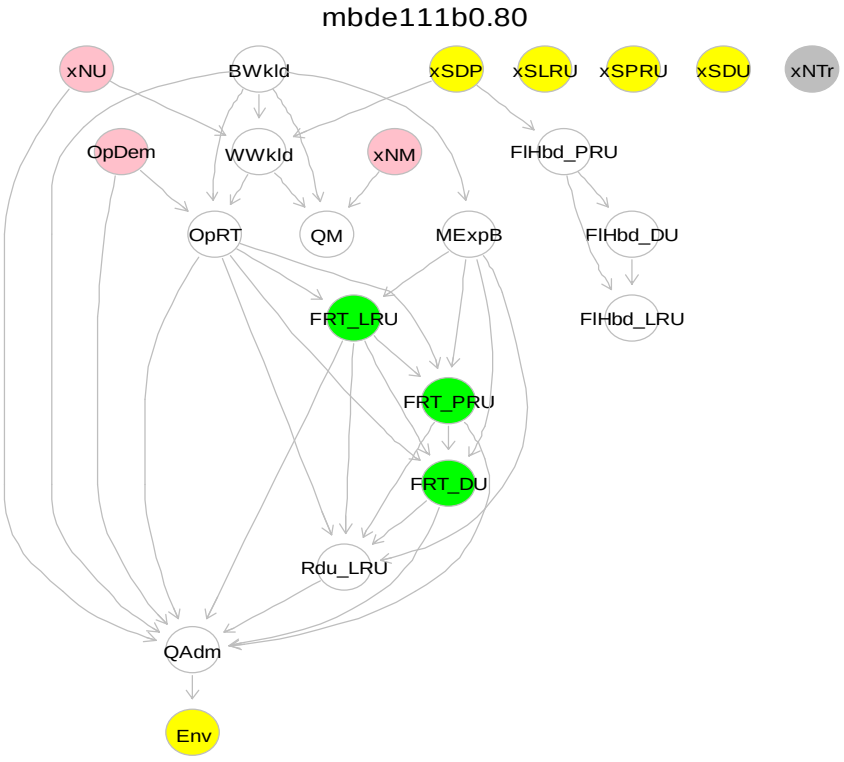


Figure A-1: All parts model, machine learnt DAG[38] (BN 5)

[38] Observe the participation of the $xSDP$

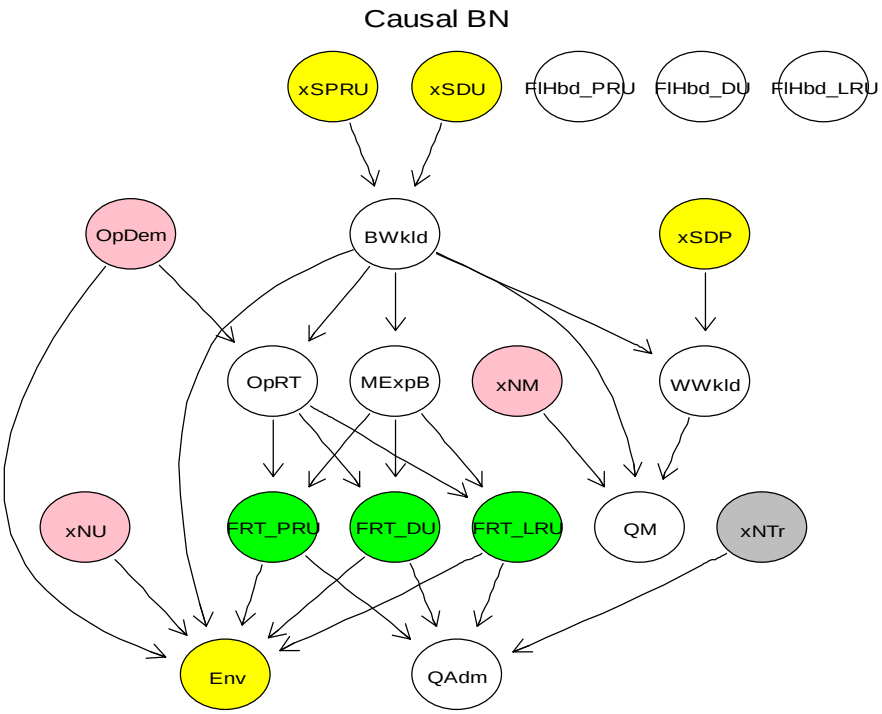


Figure A-2: All parts, elicited DAG (BN 6)

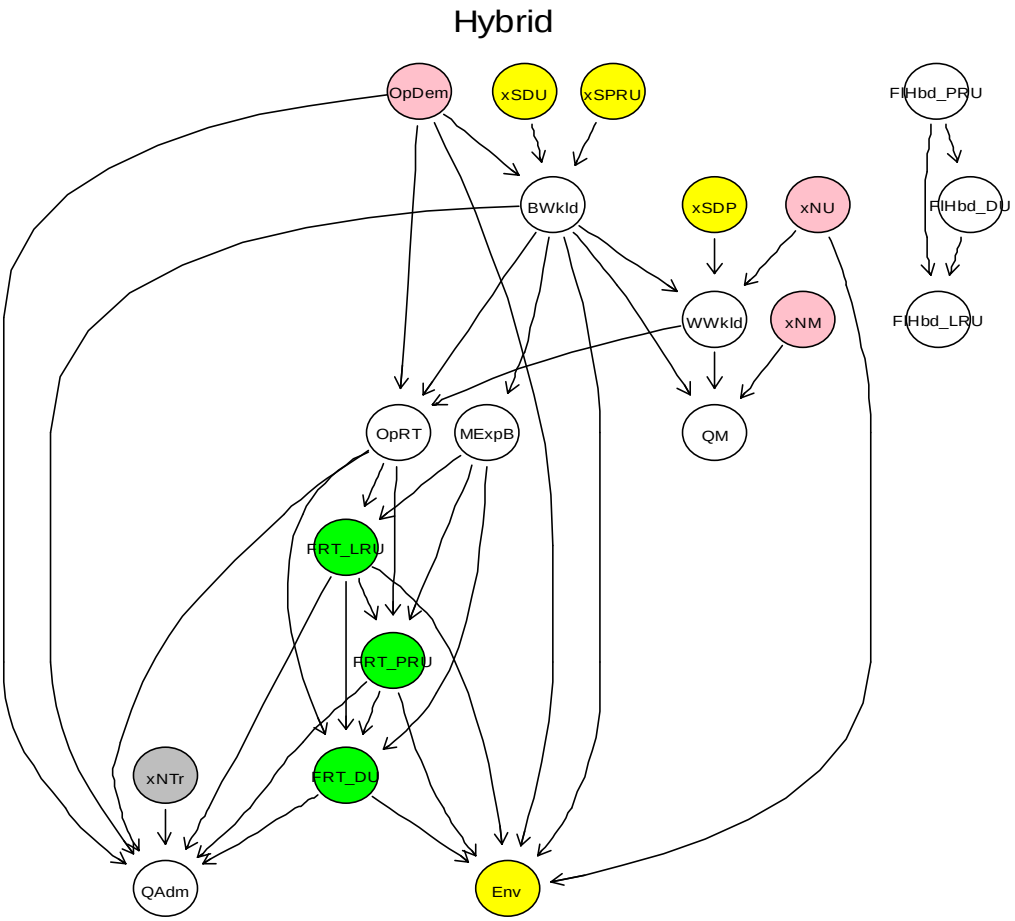


Figure A-3: All parts, hybrid DAG that maintains (parts of the) elicited[39] (BN 8)

[39] Observe the participation of the $xSDP$

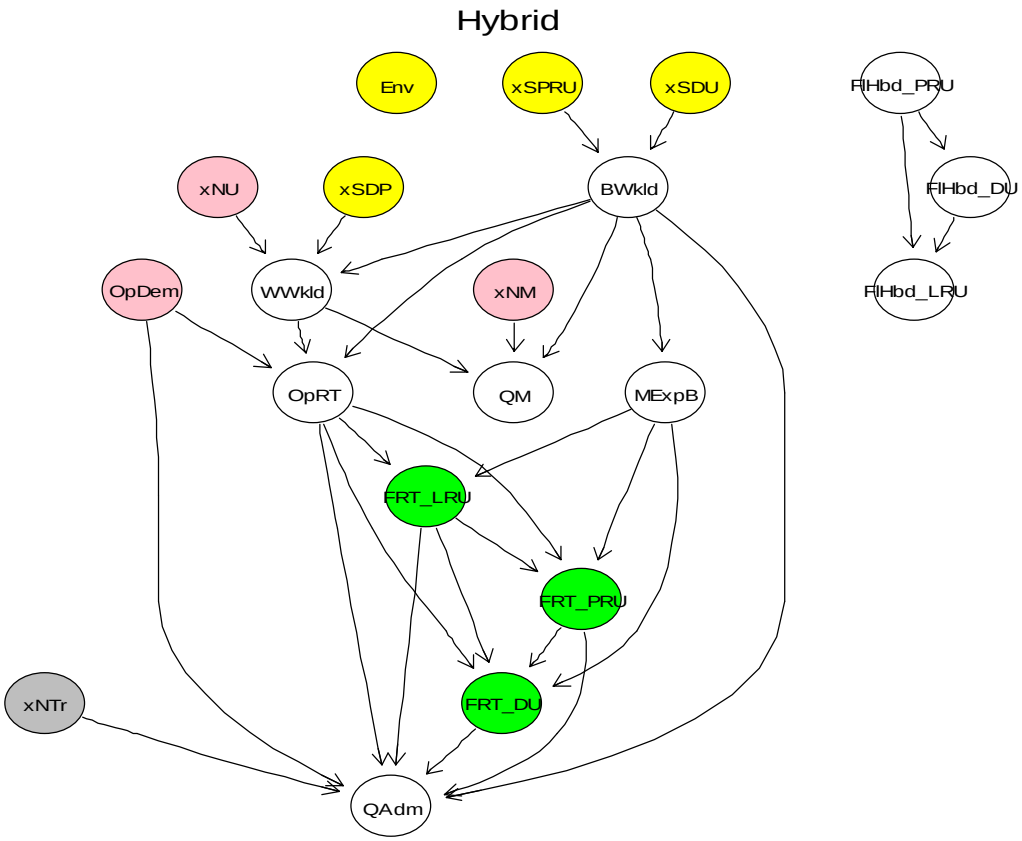


Figure A-4: All parts, hybrid DAG that started from (parts of the) elicited (BN 7)

A.1.2 A BN DAG for Only the FRT of the LRU

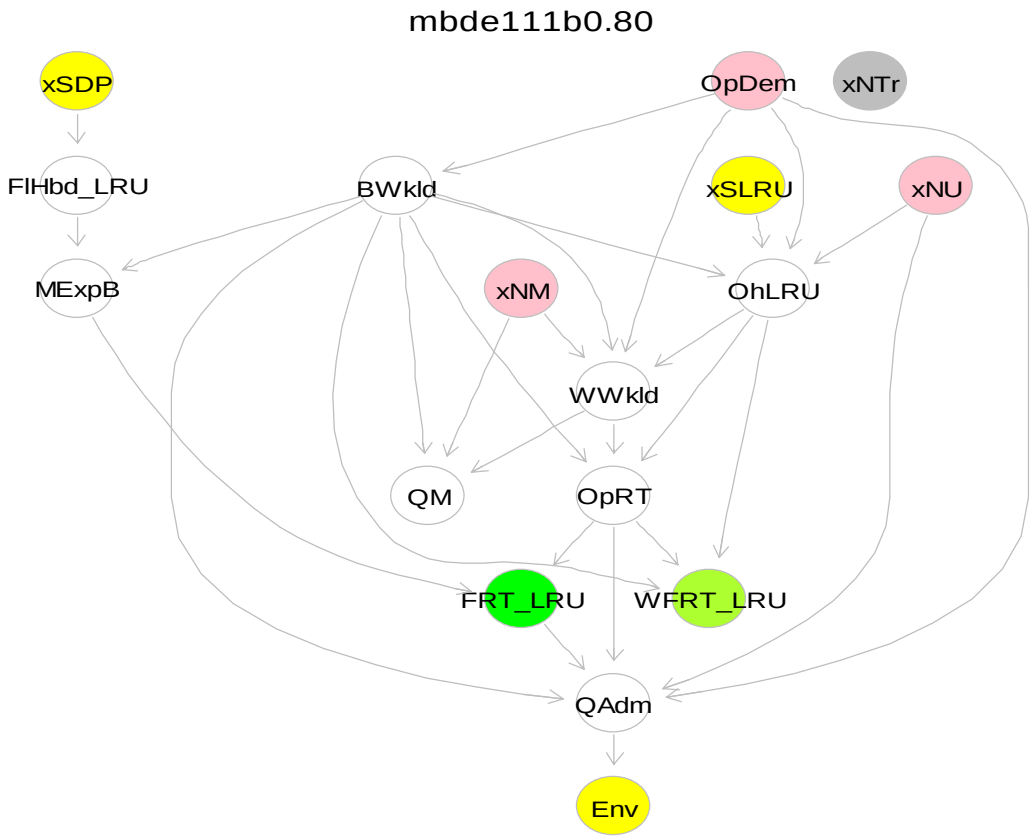


Figure A-5: LRU only, machine learnt DAG (BN 1)

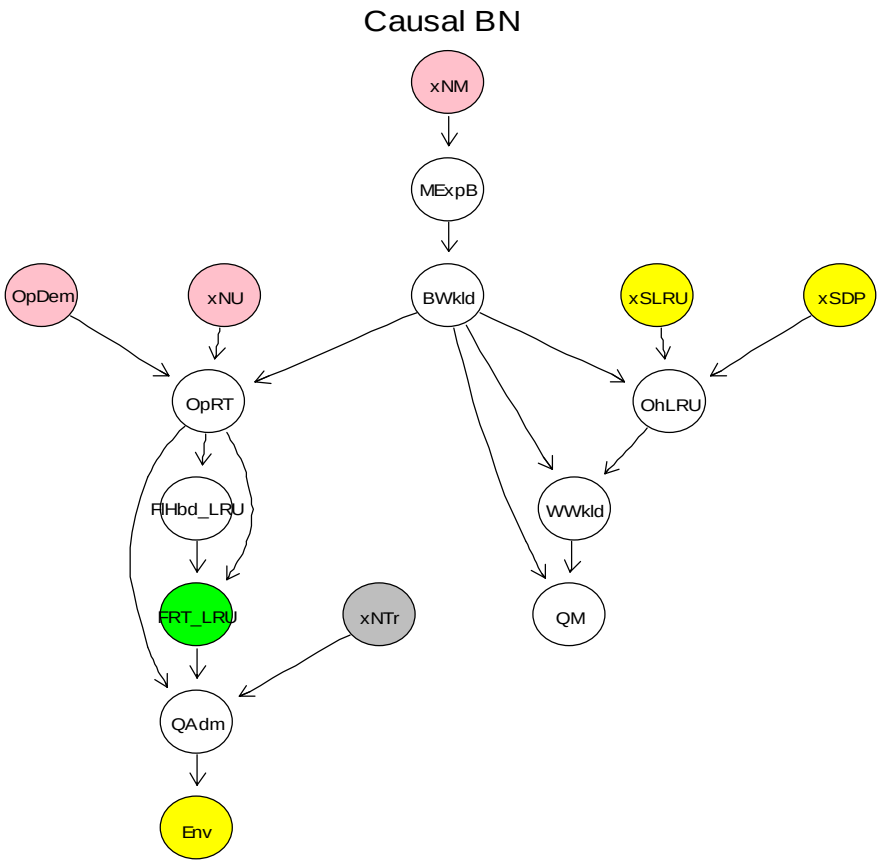


Figure A-6: LRU only, elicited DAG (BN 2)

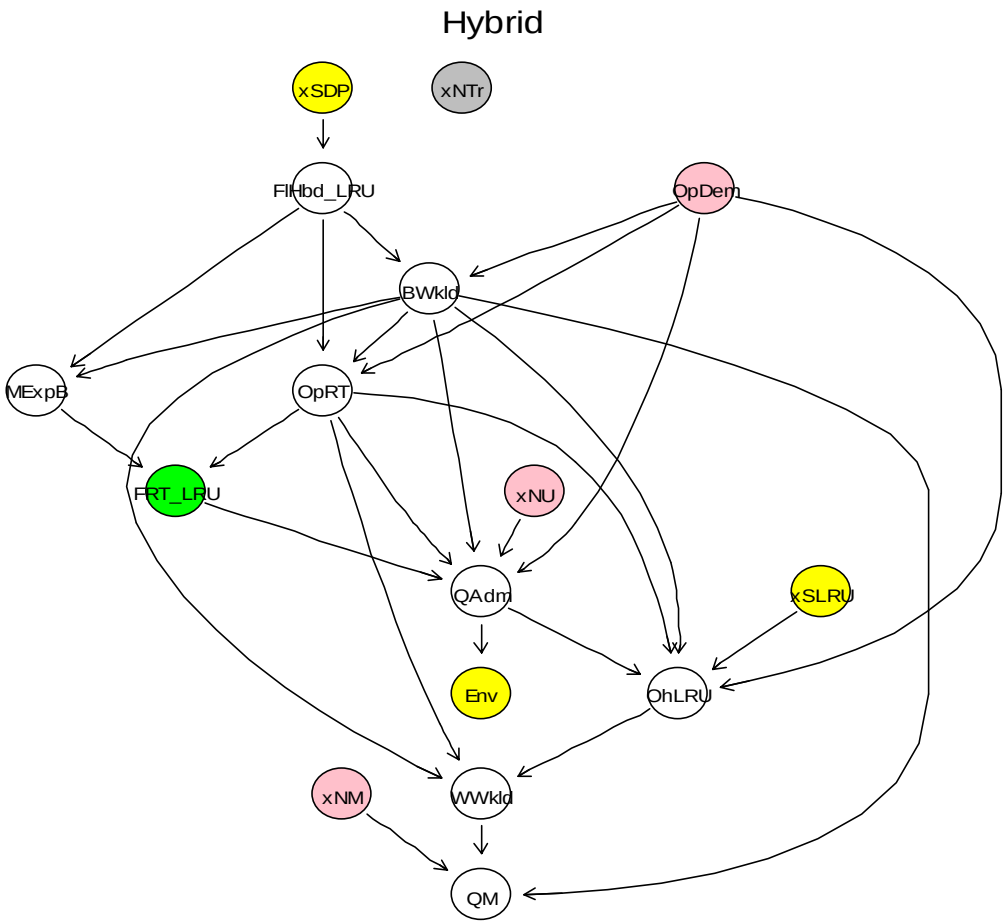


Figure A-7: LRU only, hybrid DAG that maintains (parts of the) elicited (BN 4)

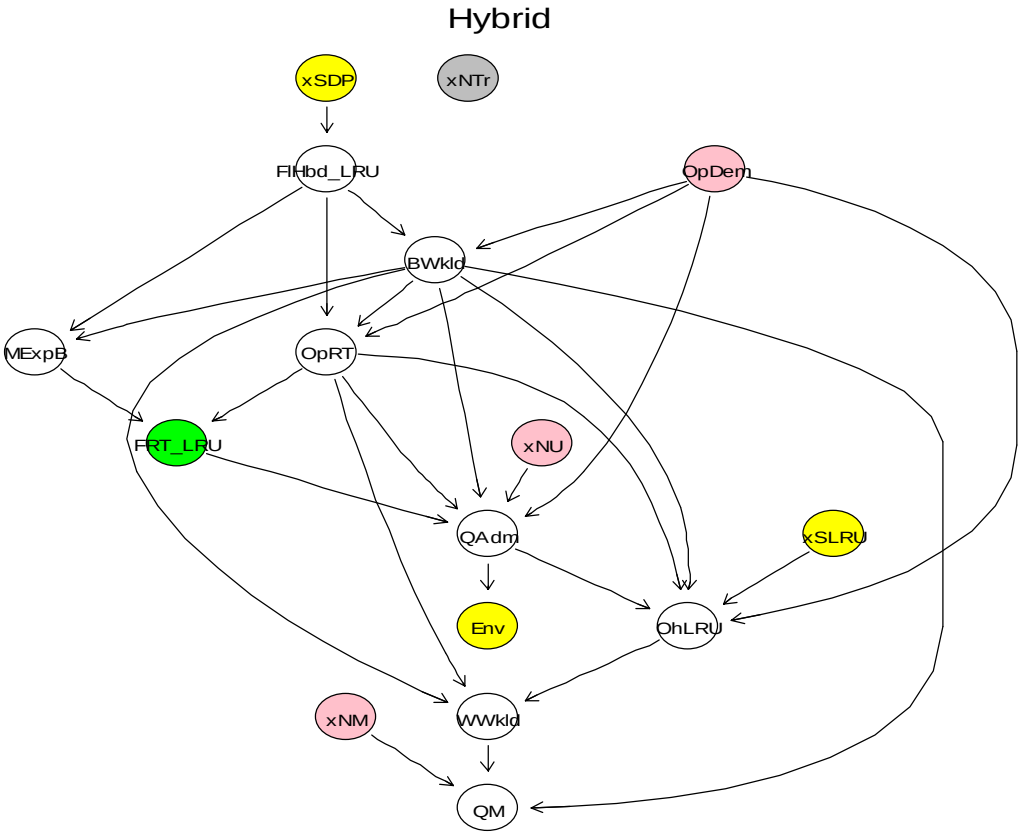


Figure A-8: LRU only, hybrid DAG that started from (parts of the) elicited (BN 3)

A.1.3 A Logistic Regression Model for Only the FRT of the LRU

$$logit(FRT_{LRU}) = b_0 + b_1 OpDem5/5{:}xSDP + b_2\, EnvNotOK{:}xSDP + b_3\, xSDP$$

The coefficients of b_0, b_1, b_2 and b_3 are -4.71342, -0.03432, 0.03321, and 0.12309 respectively, with standard errors of 0.34858, 0.03141, 0.02482 and 0.08373. The reference settings of the variables are '4/5 of a day' for $OpDem$, 'OK' for Env and "No failure" for the FRT.

None of the coefficients is significant at the 5% level but the cross validation tests showed that this model gave the best prediction out of the ones tested. What the model's coefficients show is that for any of the values of the $xSDP$, the log odds of getting a failure is increased when the Operational Demand increases (sum of b_1 and b_3) or when the Environmental conditions get worse (sum of b_2 and b_3)

In order to forecast demand for Phase 9, where the state of the Env variable is not yet known but there is a probability distribution for it, the forecast uses the

probability values as weights for a weighted average of the two outputs obtained using the two possible values for the Environment.

A.2 Models for the Forecast of the Demands in PRU

A.2.1 A Single BN with All the FRT Nodes Included

See Figure A-1, Figure A-2, Figure A-3 and Figure A-4 above.

A.2.2 A BN DAG for Only the FRT of the PRU

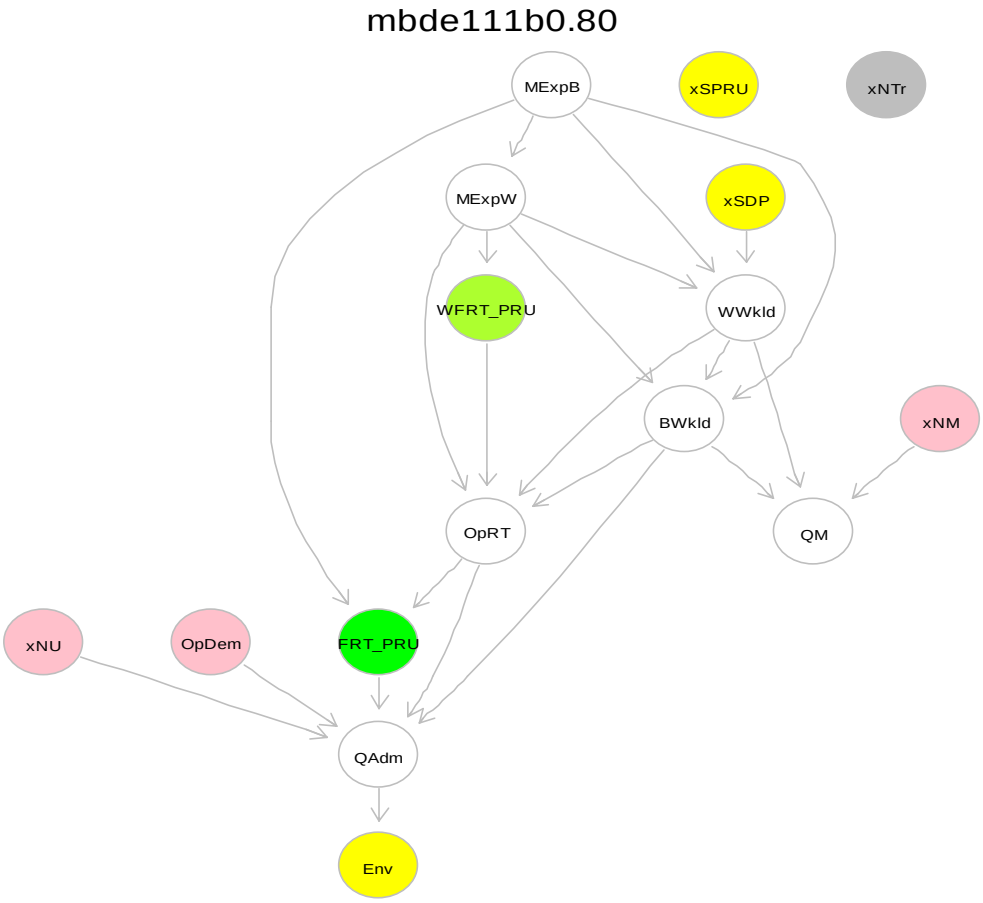


Figure A-9: PRU only, machine learnt DAG[40] (BN 1)

[40] Observe the participation of the $xSDP$

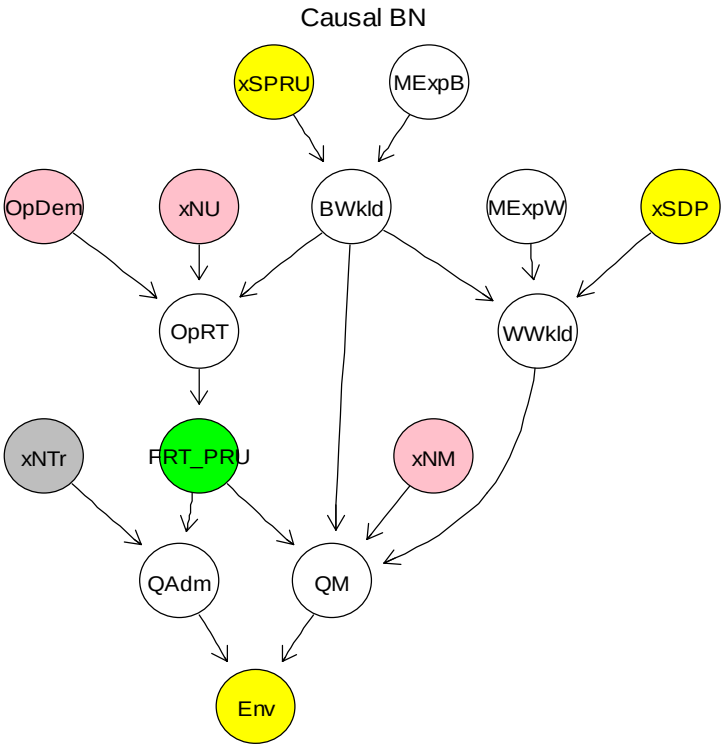


Figure A-10: PRU only, elicited DAG

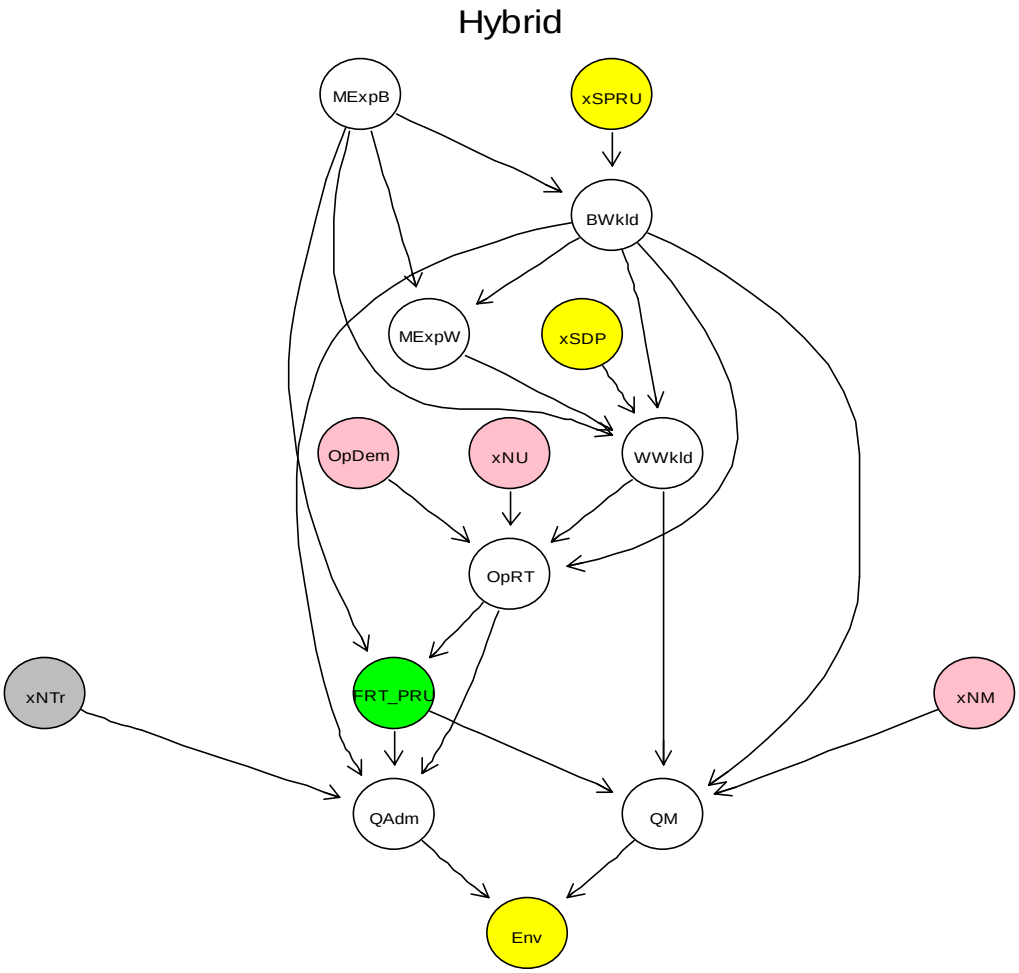


Figure A-11: PRU only, hybrid DAG that maintains (parts of the) elicited[41] (BN 4)

[41] Observe the participation of the $xSDP$

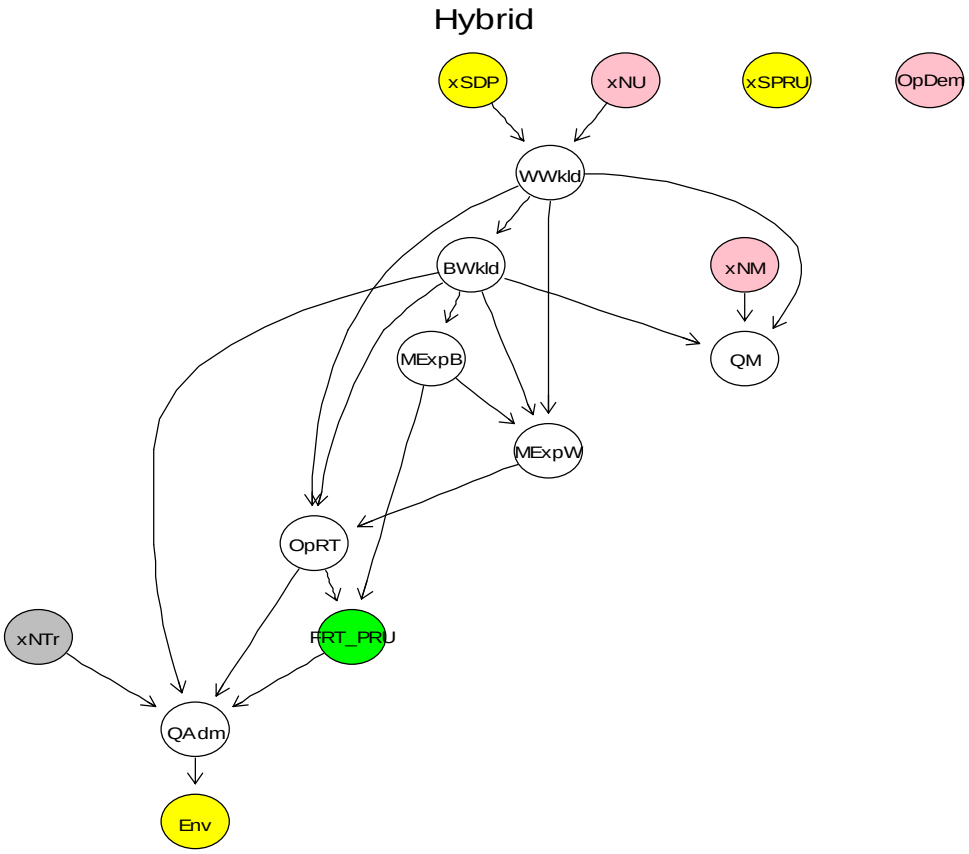


Figure A-12: PRU only, hybrid DAG that started from (parts of the) elicited[42] (BN 3)

A.2.3 A Logistic Regression Model for Only the FRT of the PRU

$$logit(FRT_{PRU}) = b_0 + b_1 OpDem5/5 + b_2\, xSDP + b_3\, xNM + b_4\, xNTr$$

The coefficients of b_0, b_1, b_2, b_3 and b_4 are -5.97291, 0.61480, 0.05437, 0.54505 and -0.22579 respectively, with standard errors of 0.64786, 0.28305, 0.11968, 0.36318 and 0.25177. The reference settings of the variables are '4/5 of a day' for $OpDem$ and "No failure" for the FRT.

Only the $OpDem$ coefficient is significant at the 5% level but the cross validation tests showed that this model gave the best prediction out of the ones tested. What the model's coefficients show is that as the values of the all the predictors apart

[42] Observe the participation of the $xSDP$

from $xNTr$ increase (or the $OpDem$ changes to 5/5), the log odds of getting a failure increases as well.

A.3 Models for the Forecast of the Demands in DU

A.3.1 A Single BN with All the FRT Nodes Included

See Figure A-1, Figure A-2, Figure A-3 and Figure A-4 above.

A.3.2 A BN DAG for Only the FRT of the DU

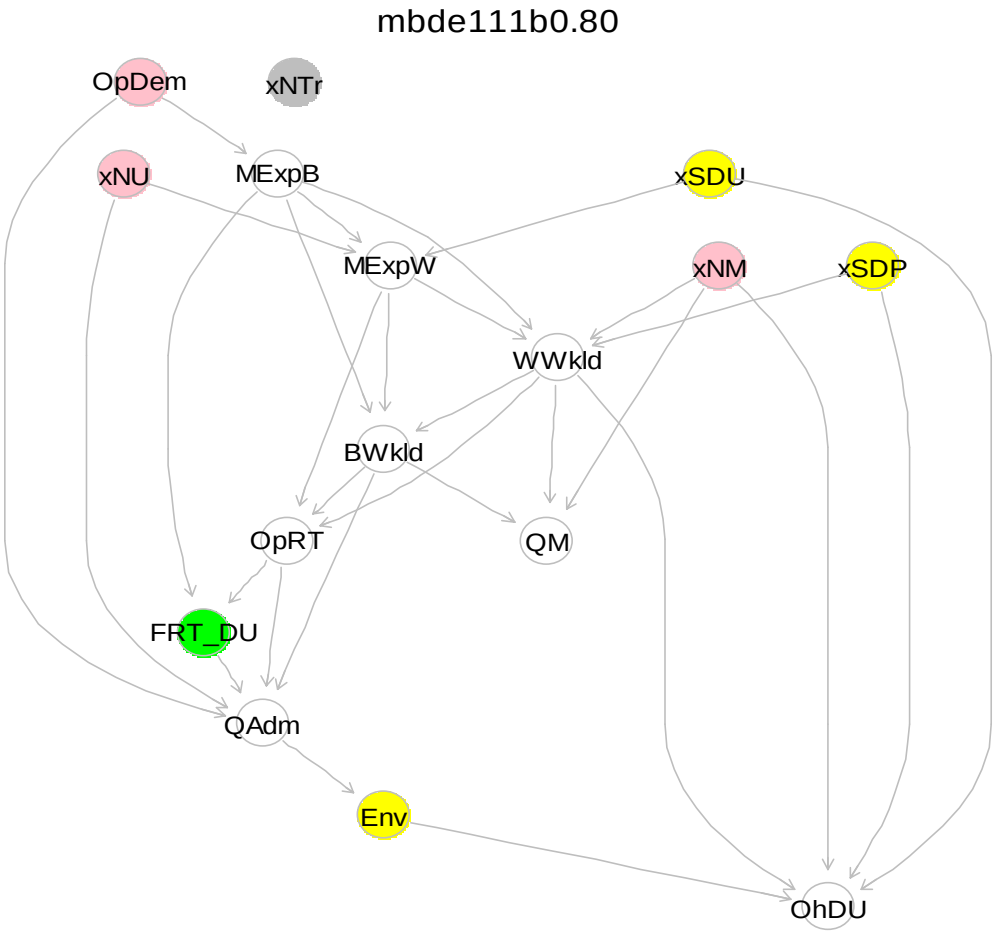


Figure A-13: DU only, machine learnt DAG[43] (BN 1)

[43] Observe the participation of the $xSDP$

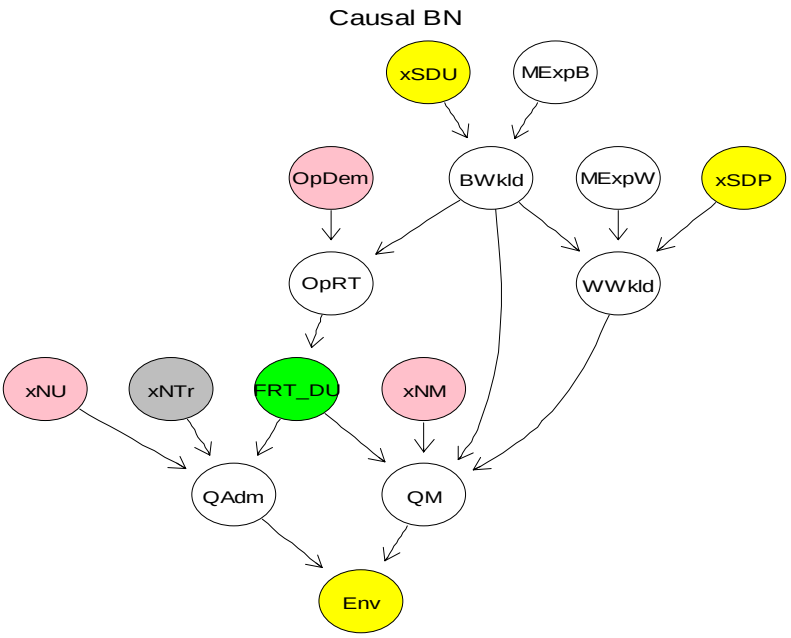


Figure A-14: DU only, elicited DAG (BN 2)

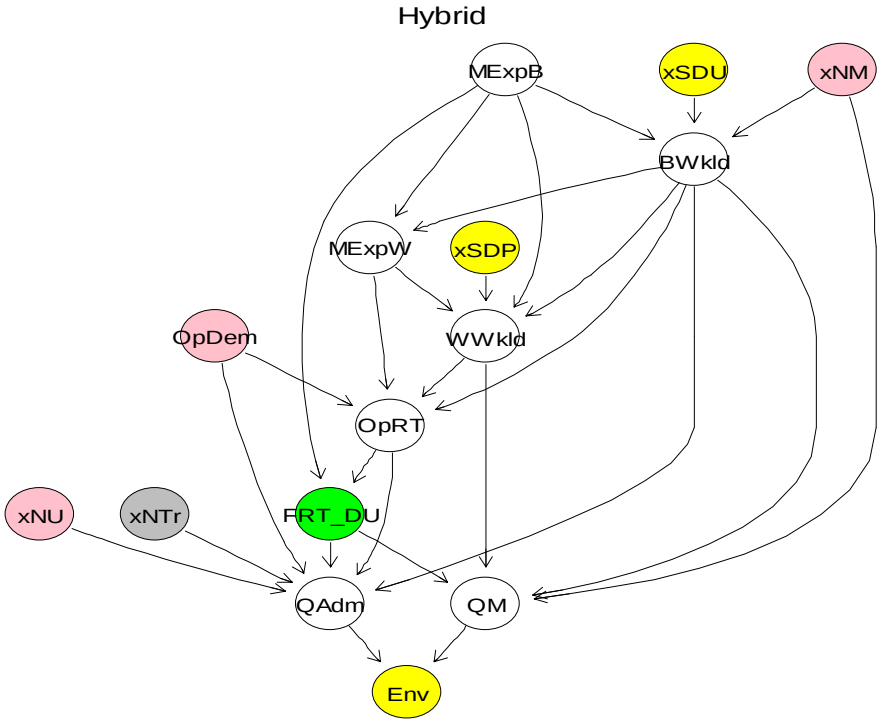


Figure A-15: DU only, hybrid DAG that maintains (parts of the) elicited[44] (BN 4)

[44] Observe the participation of the $xSDP$

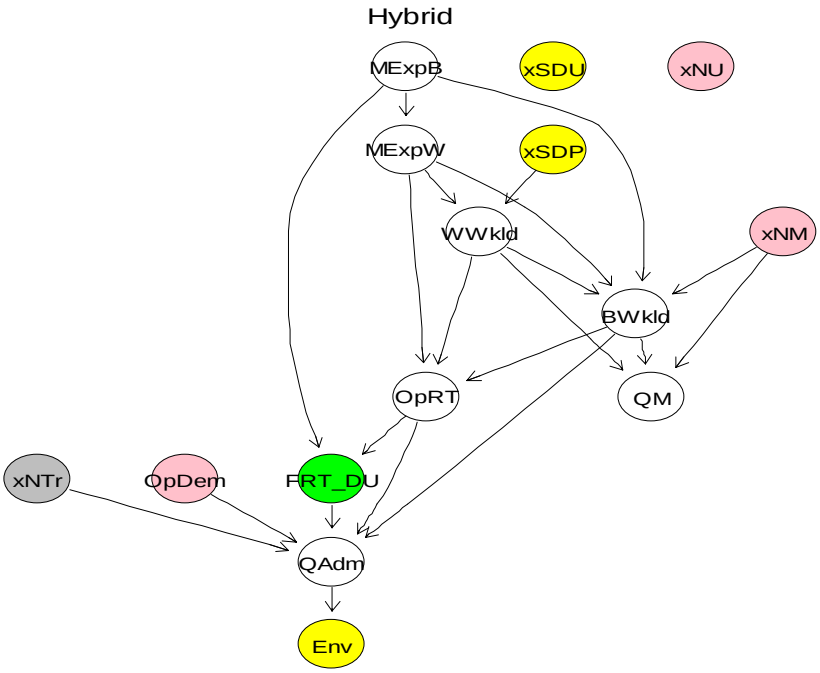


Figure A-16: DU only, hybrid DAG that started from (parts of the) elicited[45] (BN 3)

A.3.3 A Logistic Regression Model for Only the FRT of the DU

$$logit(FRT_{DU}) = b_0 + b_1 OpDem5/5 + b_2 xSDP + b_3\, EnvNotOK + b_4\, OpDem5/5{:}\, xSDP$$

The coefficients of b_0, b_1, b_2, b_3 and b_4 are -4.879939, 0.712373 , 0.004159, 0.223769 and -0.078618 respectively, with standard errors of 0.48648, 1.287017, 0.115224, 0.158063 and 0.240377. The reference settings of the variables are '4/5 of a day' for $OpDem$, 'OK' for Env and "No failure" for the FRT.

None of the coefficients is significant at the 5% level but the cross validation tests showed that this model gave the best prediction out of the ones tested. What the model's coefficients show is that for any of the values of the $xSDP$, the log odds

[45] Observe the participation of the $xSDP$

of getting a failure is increased when the Operational Demand increases (sum of b_1 and b_4) or when the Environmental conditions get worse.

In order to forecast demand for Phase 9, where the state of the Env variable is not yet known but there is a probability distribution for it, the forecast uses the probability values as weights for a weighted average of the two outputs obtained using the two possible values for the Environment.

Appendix B Observations from Contrasting the Phase 9 Outputs of the Scenarios

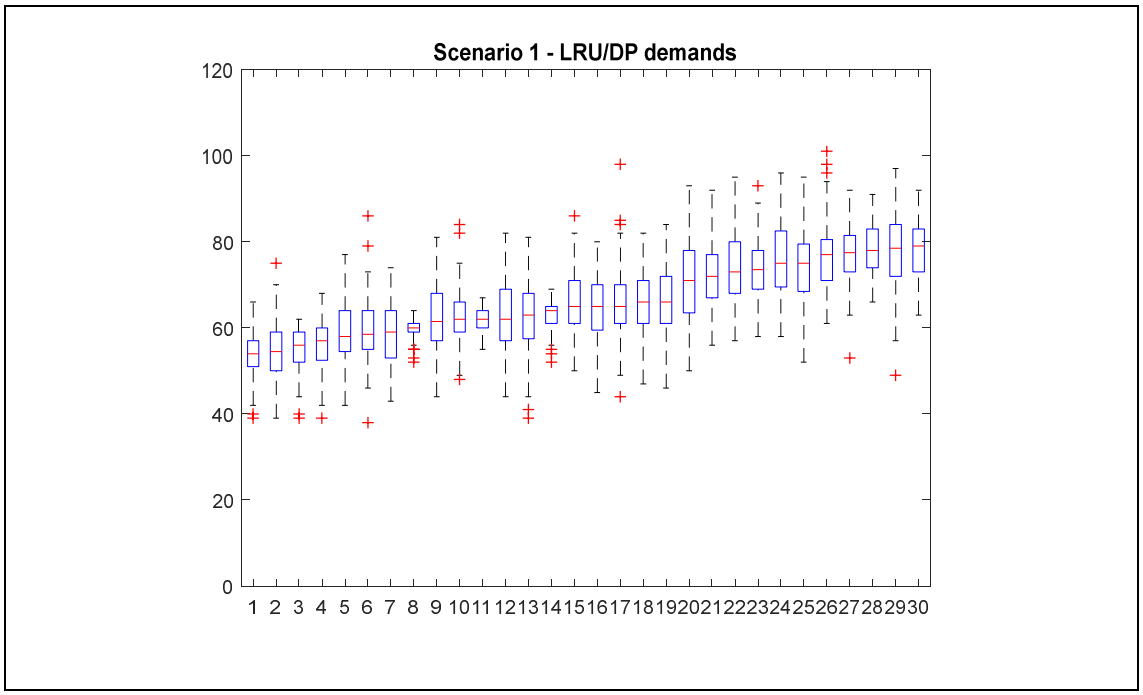
B.1 Benefits from Using the BNs to Identify the Influential Contributors

Of the two simulation scenarios cases, the main difference implemented was that the system in Case 2 (Section 7.3) had a more elaborate Equipment Breakdown Structure (EBS) than the first. The additional components that were included in Case 2 were built to be comparatively a little more reliable than the one that was used in both cases. In the current section some of the effects are explored that the changed system had in the number of the demands and an effort is also made to investigate their causes in order to inform modelling.

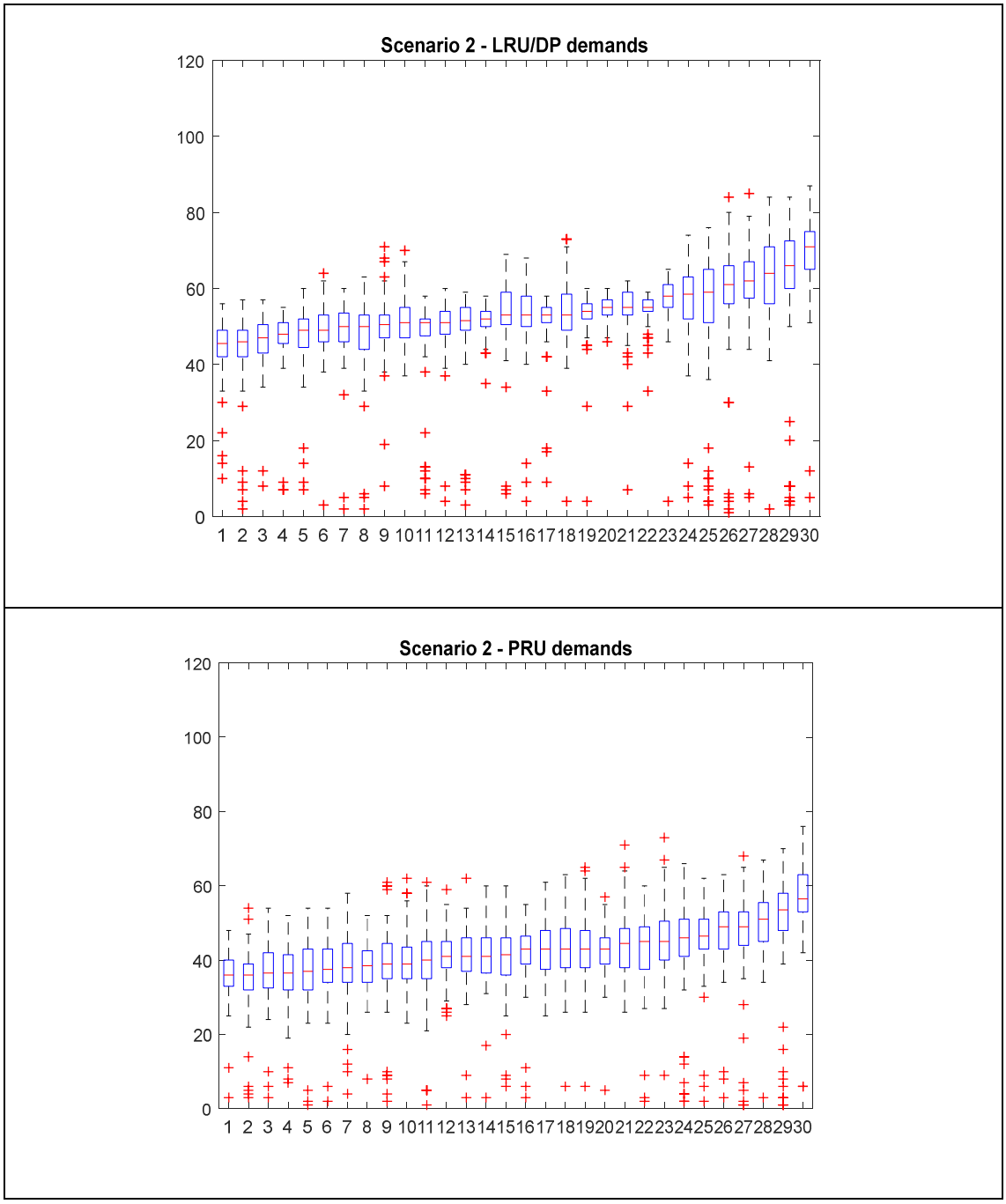
Figure B-1 shows a random sample of 30 cases from the 144 simulated in Case 1 and an equal random sample from the 512 simulated in Case 2. In all cases the boxplots have been sorted by their increasing medians. Each box includes the 25% up to 75% of the 100 replications of each test case (the 25th ($q1$) and 75th ($q3$) percentiles respectively). The red line inside the box signifies the median, while the crosses outside the box show any outlying value. An outlying value in these cases are defined as those values that are higher than $q3 + w \times (q3 - q1)$ or less than $q1 - w \times (q3 - 1)$, where w is the maximum whisker length. The default value for whisker corresponds to approximately +/–2.7σ (MATLAB, 2017). All figures have the same horizontal and vertical axes range.

The three bottom plots are from Case 2. These plots verify the fact that the DU component was slightly more reliable than the PRU which in its turn was more reliable than the LRU/DP[46] component.

The first interesting observation though comes from comparing the top two plots which are of the same component (LRU/DP). In Case 2, the component presents a potential intermittent behaviour, which means that not all the periods/months have a demand, while in Case 1 it does not. In Case 2 the 25%-75% percentile boxes are lower than the Case 1, they are more spread with more outliers, but most of all these outliers are more on the lower side with some of them occasionally reaching zero. This shows that without changing the reliability of the component, the fact that the EBS of the system has been made more elaborate, resulted in experiencing a change in its demand distribution model.



[46] The reader is reminded that the notation which shows two parts, i.e. LRU/DP is due to the fact that the LRU is repaired by a single DP which means that the demand for the LRU is due to the malfunction of the DP and thus their rate of demand is the same


Scenario 2 - LRU/DP demands
120
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Scenario 2 - PRU demands
120
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

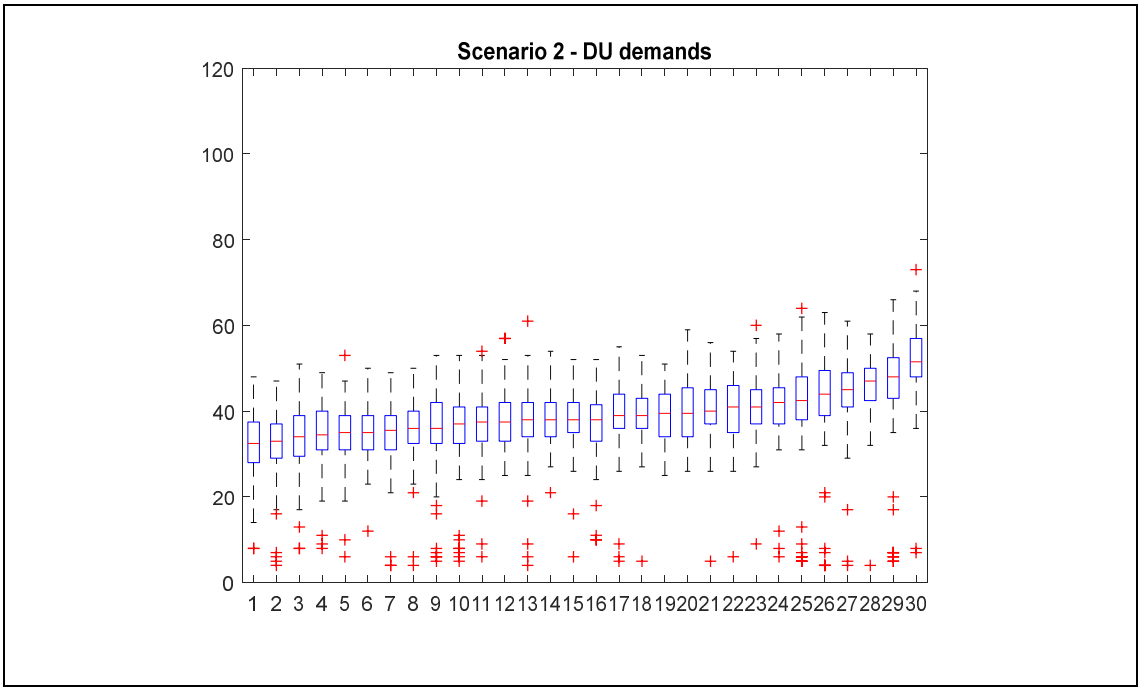


Figure B-1: Boxplots of sampled cases of Scenarios 1 and 2, sorted by their median value (four plots)

Figure B-2 presents an example of this change. The three plots on the top present the histograms of three of the 144 cases examined in Case 1 (each example replicated 100 times), while the ones at the bottom present three of the 512 cases of Scenario 2. Even though these are just three examples from each scenario, they are characteristic of the result: the shape of the demand distribution has changed in Case 2 and presents a skew to the left while it is less populated on the right side. The most interesting output of these is the fact that the component – even though as an outlier - can now present an intermittent behaviour which is a very challenging characteristic for a demand pattern[47]. Such a result can be seen in the left two time-series of Figure B-3 (each plot is a single case of the 100).

[47] Discussions on the challenges of predicting intermittent demand time-series can be found in many research papers, see e.g. Teunter and Duncan (2009), Petropoulos and Kourentzes (2015), Wallström and Segerstedt (2010), Syntetos and Boylan (2005)

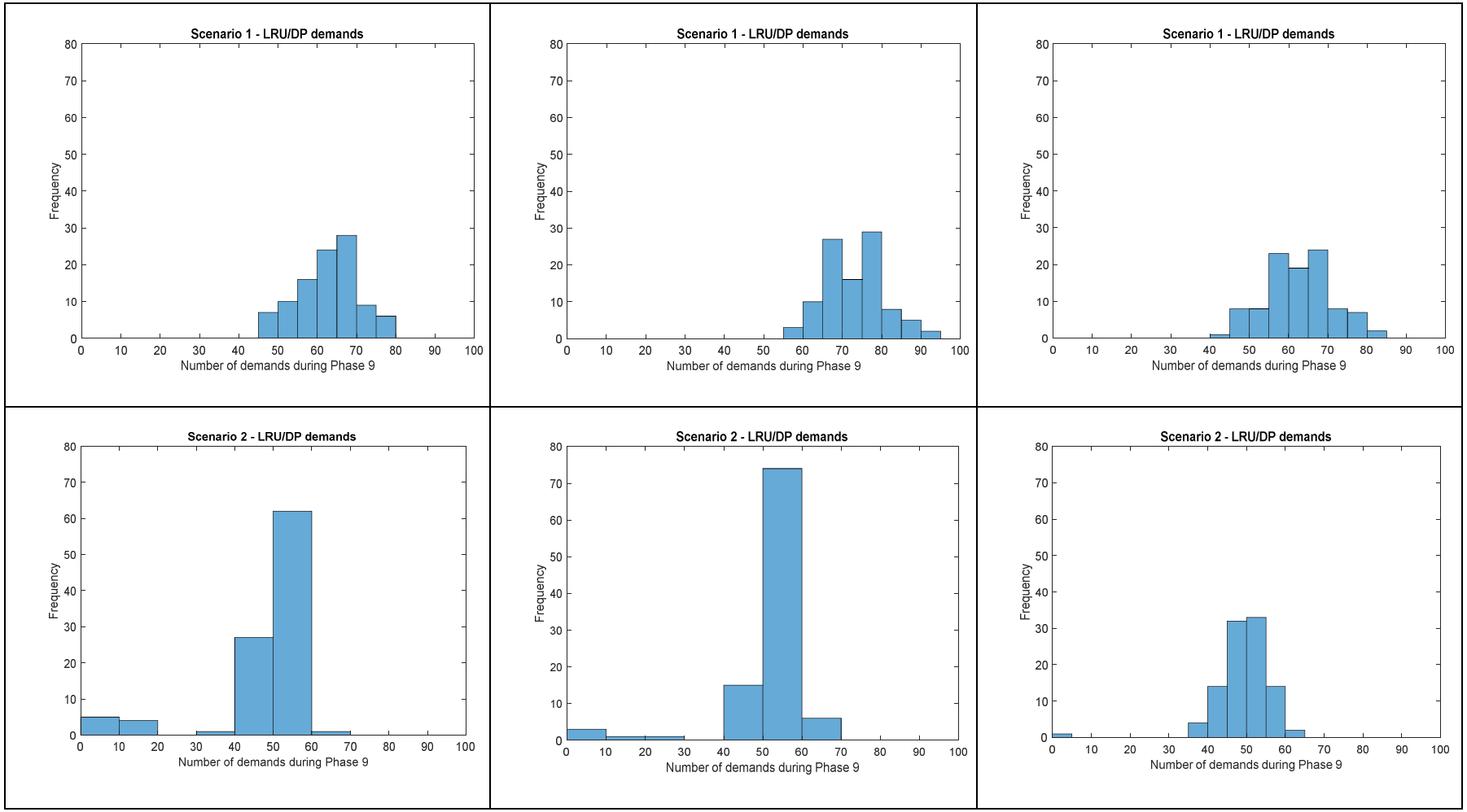


Figure B-2: Three histograms of the 144 cases (top row) vs three of the 512 cases (bottom row)

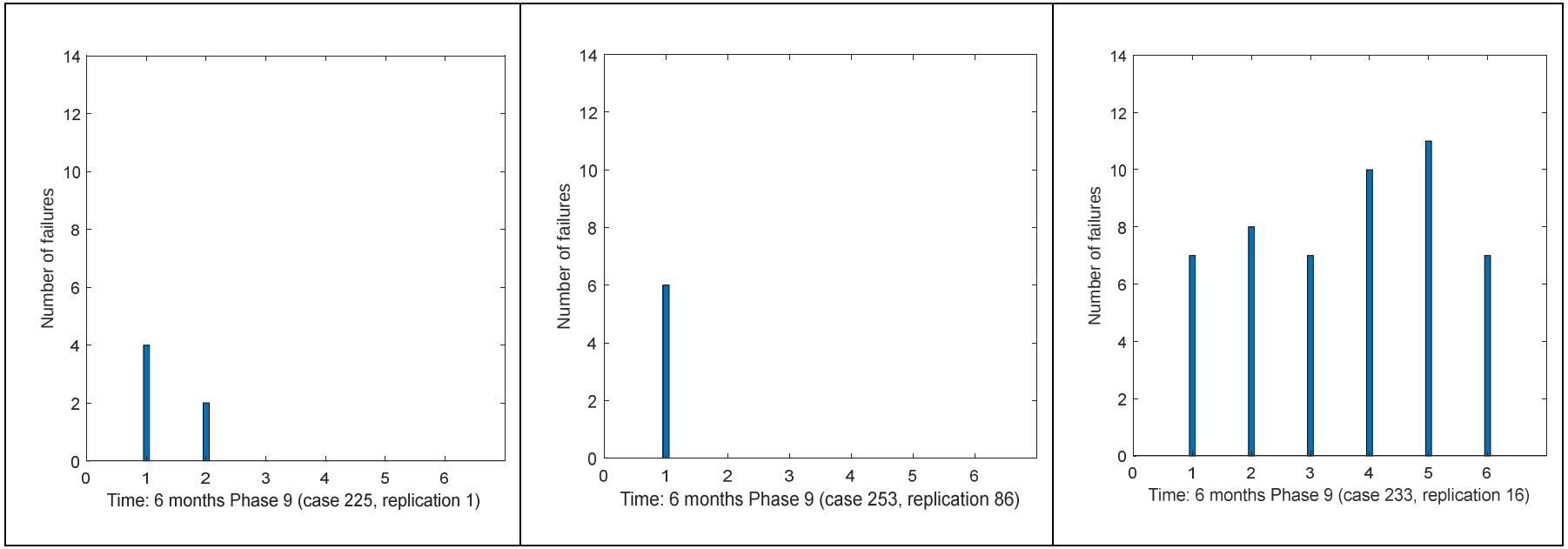


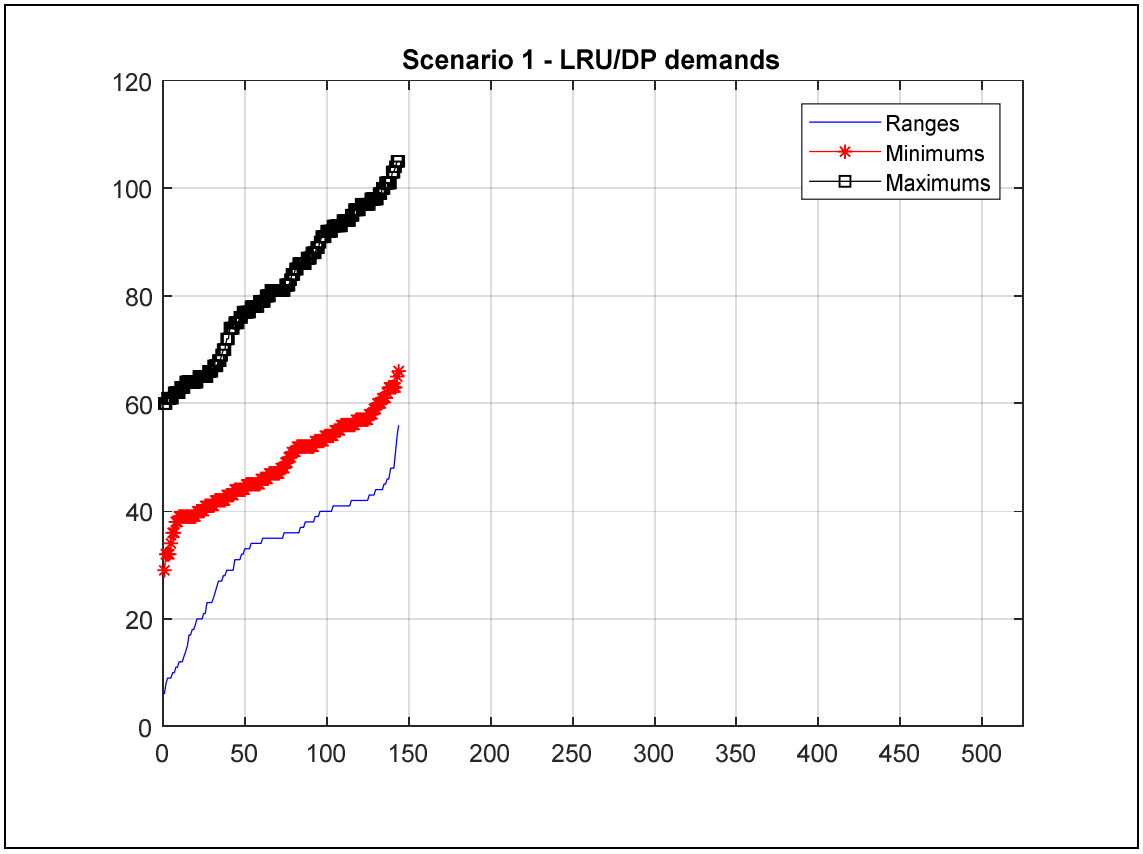
Figure B-3: Three 6-month cases out of the 512×100 that were simulated[48] [49]

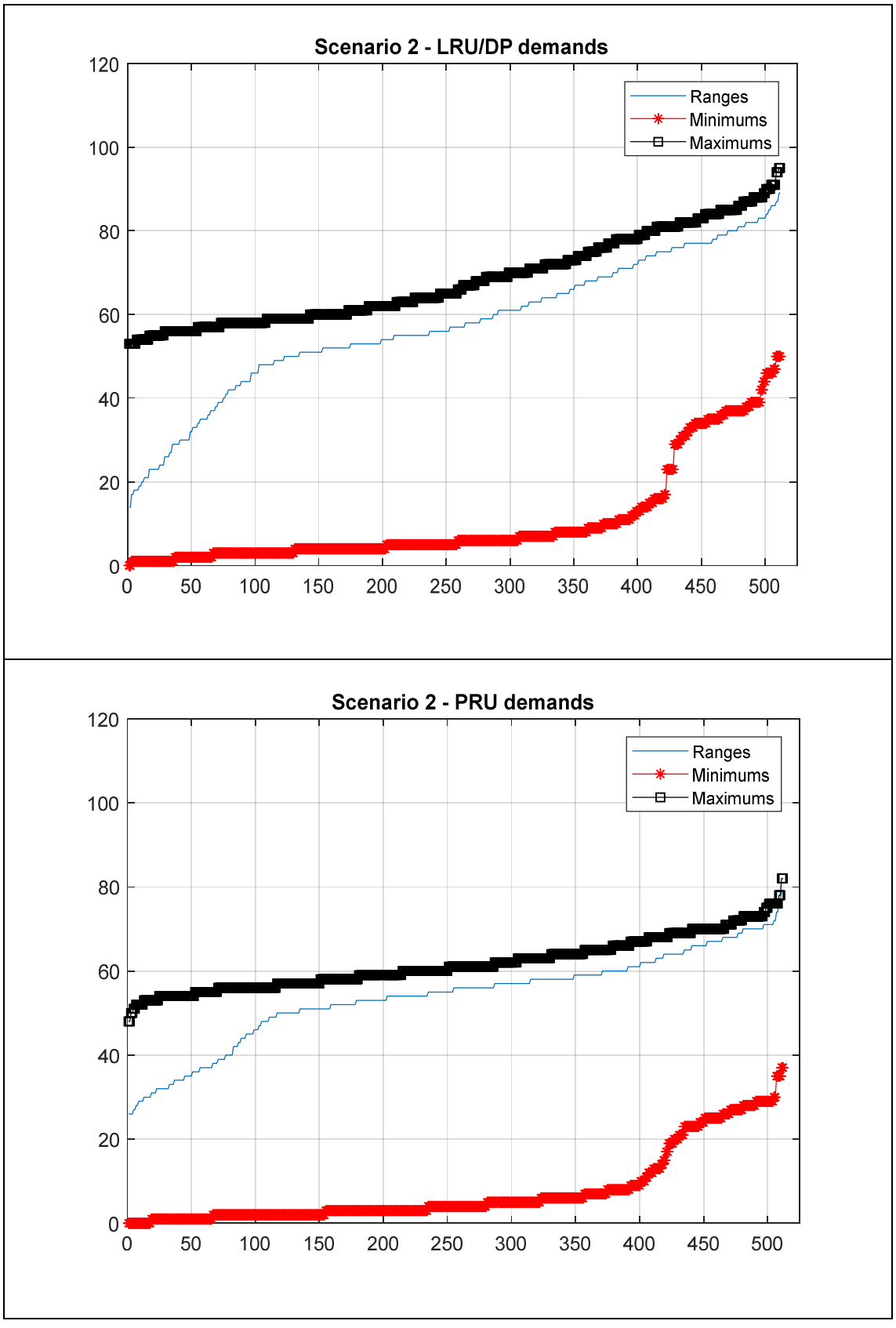
[48] Compared to the boxplots of Figure B-1, the first two time series are just two of the outliers presented as red crosses at the bottom of the boxplots

[49] The two left present a characteristic intermittent behaviour, while the third one is a common case

In order to further explore the changes in the spread of the data, Figure B-4 has also been created. In this figure, the ranges, the minimum and the maximum values of all the cases from the two scenarios for all modelled components have been plotted in an ordered sequence. The thin blue line represents the ranges, the thick red with the stars are the minimums and the thick black with the squares are the maximums.

The first thing to observe is that the line of the ranges in the first scenario is below the line of the minimums, indicating a lower spread of the data, while, in the second scenario (both for the LRU/DP and for the other two components) the line of the ranges is above the line of the minimums and can take larger values, mainly due to the fact that the minimums have dropped a lot, while the maximums have reduced but not as influentially.




Scenario 2 - LRU/DP demands
Ranges
Minimums
Maximums
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450
500
Scenario 2 - PRU demands
Ranges
Minimums
Maximums
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450
500

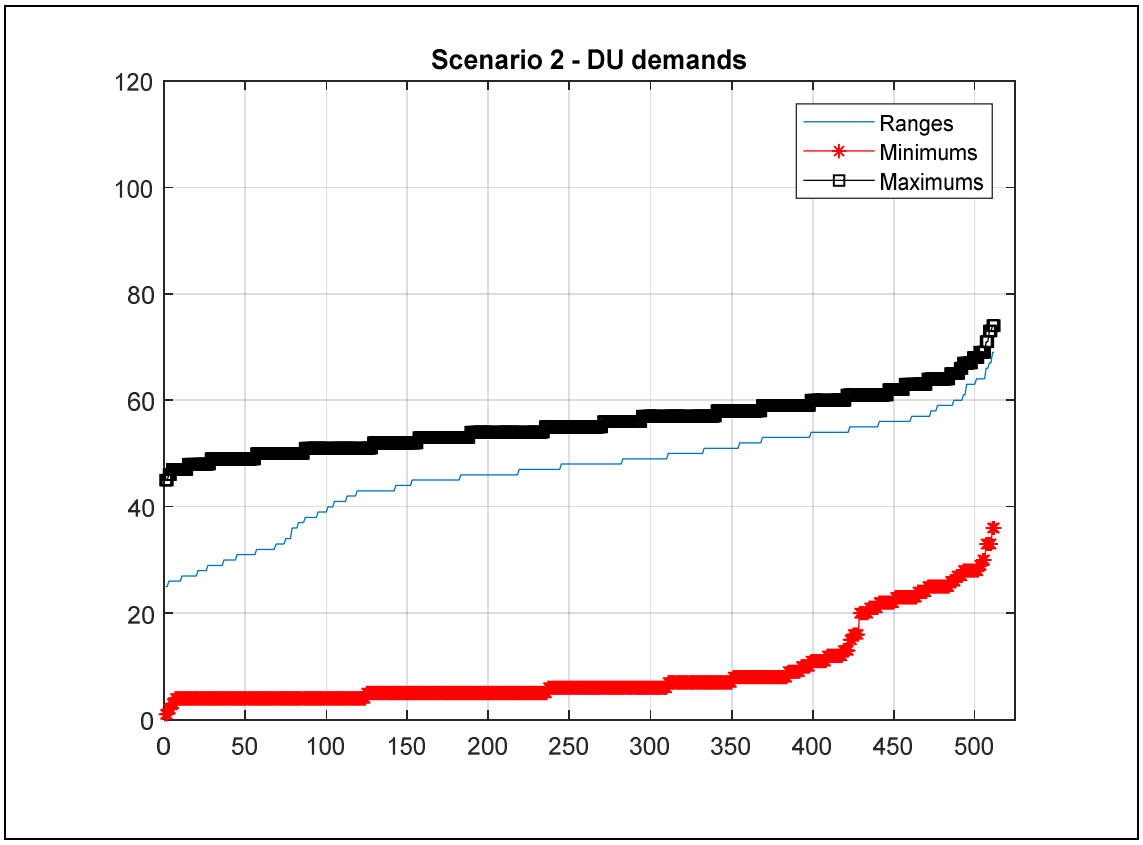


Figure B-4: Plots of the sorted values of Range, Minimum and Maximum for all values of Scenarios 1 and 2 (four plots)

The above can be also verified by the plots of the Coefficients of Variation in Figure B-5. The metric compares the standard deviation over the mean obtained from each 100-replications set, and here, it demonstrates that the second scenario produced comparatively higher spread in the values of the demand.

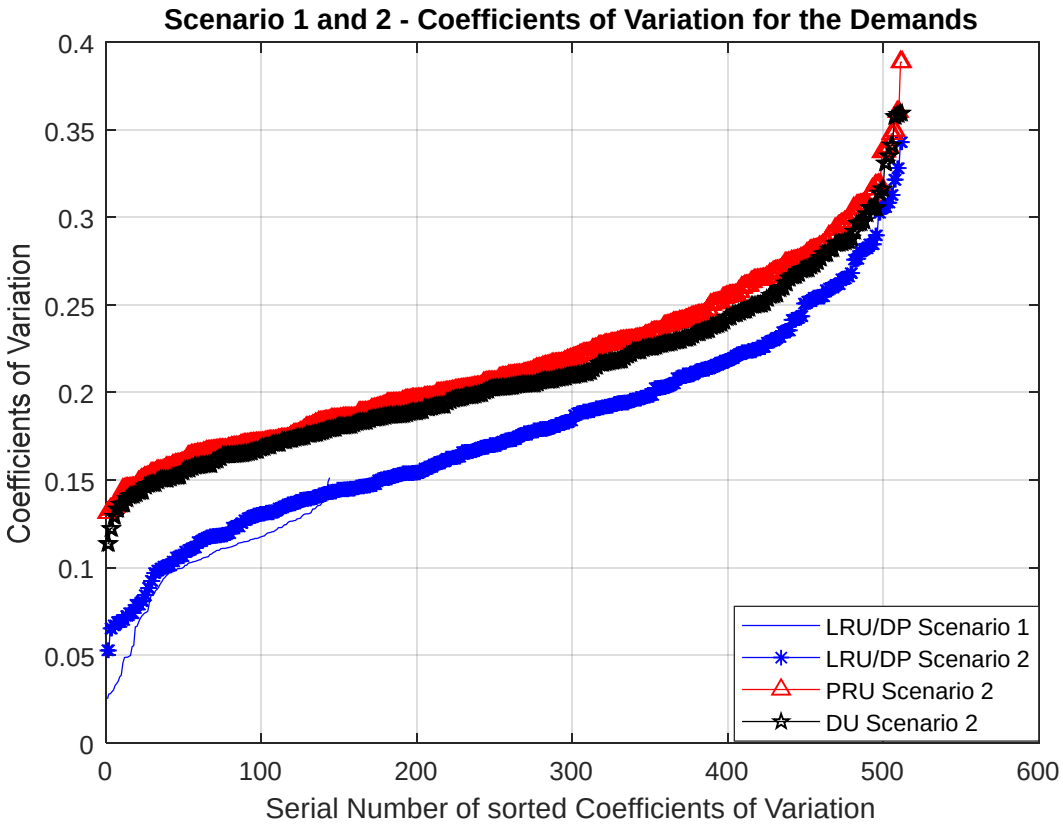


Figure B-5: Coefficients of variation for all components of both Scenarios[50]

From all the above a strong interaction can be observed among the components, and this interaction affects:

1. The location of the demand distribution, see e.g. the medians in the top two plots of Figure B-1, where the ones at the first are generally higher than the ones at the second
2. The skewness of the distribution to the left, see e.g. the increased number of outliers in the second of the top two plots of Figure B-1 and the shapes of the histograms in Figure B-2. This skewness also reaches the level of experiencing intermittency in a number of occasions
3. The spread of the distribution, see e.g. the observations about the accuracy implication metrics, where the mean root squared errors of the LRU/DP were smaller in Case 1 (Section 7.2) than the respective of Case

[50] The light line at the bottom plots the respective values of LRU/DP from Scenario 1 and the thick lines above that are from Scenario 2 for the LRU/DP, DU and PRU respectively

2 (Section 7.3) and also in Figure B-4 and Figure B-5 where the ranges and the coefficients of variation can take a lot higher values in the more elaborate EBS

What can be deduced is that the complexity of a system's EBS is one of the reasons for the above phenomena and an explanation of the mechanisms are going to be suggested in the next Section B.2. However, firstly the resulting demand intermittency is explored, which is one of the challenging situations in an SC (Fisher, Hammond, Obermeyer, and Raman, 1994).

When the BN models were being developed, it was realised that in Case 2, the $xSDP$, i.e. the nominal level of DPs in the inventory, was identified by the BN-structure machine learning algorithms as a variable that increased the predictability of all models for any of the components – i.e. regardless if it was for the LRU/DP or the PRU or the DU. Furthermore, this observation was made either when the work was during the development of the BN with all the components in a single model, or when working on the BNs for individual components. Moreover, using this observation it was also realised that the predictability of the regression models that were built, again for any of the components, increased when the $xSDP$ was included in the covariates (see Sections A.2.2 and A.3.3 for more details on the regression models of scenario Case 2). This ability of the BNs to identify factors that are not straightforward to identify is also discussed further below, after highlighting the modelling challenges resulting from comparing the examined two scenario cases.

The challenge of intermittency forecasting increases even more when one considers that the demands during Phases 1 to 8 in Case 1 and the respective of Case 2, were not very different. This can be inferred from the mean naïve forecast values of the two scenarios[51]. In Scenario 1 the mean naïve forecast was equal[52] to 3.7826, while for Scenario 2 it was:

- LRU/DP : 3.4348

[51] These are the values that were also used as denominators in the calculations of MASE
[52] Rounded to the 4th digit

- PRU : 3.0435
- DU : 2.3478

What can be observed from the above values is that the values of the respective LRU/DP outcomes are not very different in the two scenarios. This fact indicates that the initial phases produced not very different demands for this component which was common in the two scenarios. In other words, even though both scenarios had a similar number of demands in their training dataset, due to the fact that the second scenario employed a more elaborate UAV system, the likelihood of experiencing challenging very low and even intermittent demands increased, suggesting two modelling challenges.

Firstly, one would expect that models like the commonly applied time-series statistical models that do not use explanatory variables for their forecasts, could be misled and not identify the differences on the infinite-time horizon phases and the forthcoming final phase, a fact which is very important in the FPP cases where there would be only a single opportunity to provide and use a forecast.

Secondly, even more, models like regression, that do use explanatory variables could skip / miss trying to use an “irrelevant” covariate like the $xSDP$ when the model was about the demand of a different component than DP.

This observation amplifies the need to use models like the BNs that can both identify but also highlight in their graph those factors which are more influential. Consequently, the fact that the BN DAG can show which of the components are more responsible for the intermittency of the rest – the DP in the examined cases - is an important finding for the demand forecasting and consequently for the better planning as well, and this finding can be used for the development of other models as it was done with the regression models.

Finally, going back to the other observations, it can also be said that the fact that the spread and the skewness of the demand distribution is also affected by the changed EBS, indicates that these also need to be modelled using explanatory variables, something that was not addressed in the current research.

B.2 Explaining the Intermittent Behaviour of the Components

Two reasons have been identified for which the LRU/DP was actually responsible for the observations above. Firstly, the LRU/DP was the least reliable of the components. Due to this fact, the SC system would experience more failures of the LRU/DP than the rest. Therefore, the LRU/DPs requirement for repair/resupply resources was the reason for any of the other components to "extend" their calendar life by waiting for the execution of the repair and resupply activities for the LRU/DP. Consequently, this extension in the other parts' calendar life resulted in experiencing a reduction in their failure rate as well. On the other hand, this also explains why the LRU/DP demand pattern presented intermittent behaviour in the second scenario: its own calendar life was "extended" due to the presence of the other components and their breaks, repairs and logistic delays.

Secondly, apart from being the least reliable, the LRUs' repair process is longer and requires more resources. So, by the time an LRU would require repair, apart from the common first-line diagnosis and repair delays on the UAV, there was also a second line delay due to the diagnosis and repair at the shop, something that was not included at such a length for the other components and thus made the LRU/DP even more influential.

Regardless of the mechanism that is behind these phenomena, from a practical perspective what needs to be again stressed is that a model like the BN can alert the modeller of the existence of such phenomena and also of the factors which are more influential.

Using what was observed from the case studies and the conceptual models that were introduced in Chapter 3, in what follows, the factors that might control the calendar life and thus the experienced failure rates of the components are generalised.

The time from when a component is put into operation until it breaks down can be partitioned into the following two parts:

1. Its wait duration which includes any kind of non-operational time while it is mounted on the system of which it is a component. For those systems that

are on continuous operation, e.g. a power generator, such a time includes the waiting time for which the system is under maintenance due to a different component

2. Its actual working / operational duration which is related to its reliability

Placing each of the above two categories under the lenses of the individual contexts that were identified as a conceptual model in Section 3.3 (Engineering, Environment, Operations and Support), one can see the way that they work on the total duration of a component's life without repairs - its calendar life - and consequently the rate that breakdowns are experienced and consequently demand for spares as well.

On the one hand, there is the wait duration (the first of the two categories in the list above) inside the system on which the component has been mounted. As it was mentioned earlier, the important observation in this case, is that this duration is not affected necessarily by the effects of the contexts on this specific component. This duration is affected by the stand-by operational duration ("not-used but working"), but also by the effects of the previously mentioned contexts on the specific components that are mounted on the same system, but which are more prone to repair and logistic delays and which can be identified by models like the BNs. This is part of the Engineering context perspective and it includes both those influential components' Reliability and Maintainability. Moreover, in order to explore and model the wait duration of any component, the Environmental, Operations and Support contexts' effects should be considered on those influential components since it is their durations that drive the wait durations of the other components.

On the other hand, the operational duration /reliability of a component (the second of the two categories in the list above), apart from its nominal engineered life, it is also affected by the Environmental conditions under which it operates, the proficiency of its operators and the Operational/usage rate, and the quality of its

Support/repair. This set of factors can extend or reduce the component's operational duration.

In the next Section B.3, the explanations provided above have been used to further extend the previous observations into the examination of the factors that compose the Operational Availability function.

B.3 Looking Closer at the Operational Availability Function

If the LRU/DP component is seen not just as a part but as a repairable subsystem - a system with its own functional output that contributes to the whole system's Operational Availability – the previous observations can be extended and probably (re)form the thinking when the Availability of that subsystem is calculated.

The Operational Availability function is used either to evaluate the performance of a system, or even plan its availability as a requirement:

$$A_o = \frac{MTBM}{MTBM + MTTR + MLADT}$$

$MTBM$: Mean Time Between Maintenance activities (either corrective or preventive)

$MTTR$: Mean Time To Repair

$MLADT$: Mean Logistics and Administrative Delay Time

Firstly, the two observations that were discussed in Section B.2 are directly related to the function's numerator $MTBM$. Both the reliability, but also the waiting periods of the component as a subsystem are included in the $MTBM$. However, the waiting periods might not be driven by the subsystem itself but by one or more other subsystems (components), possibly irrelevant to the one under consideration, and which are the ones that are more responsible than the rest for the range of values of the logistic delays which affect the whole (super)system that they belong.

Moreover, the discussed effects on the $MTBM$ can also be related to subsystems that are planned to work continuously (e.g. the power unit of a critical system at

a hospital) and a waiting duration might be not thought of being included in their A_o function. Again, the numerator of their availability function might be better approximated if it includes the wait time due to other influential parts.

Furthermore, the denominator of the A_o is also affected by components different to the one considered by the A_o function through its $MLADT$ term. Again, one would expect that the term would be more effectively calculated if one considers not only the subsystem/component under consideration but also others that are more influential. In this case, the subsystem/component under consideration might be waiting on the workbench to be repaired not because a spare related to its maintenance is absent, but because another part, more “sensitive”, is under repair.

Finally, these observations are also relevant when the (super)system undergoes a modification, since even if this modification is irrelevant to some of the subsystems, it can still affect their availability calculations.

Appendix C Pre-print of “Using Bayesian Networks to Forecast Spares Demand from Equipment Failures in a Changing Service Logistics Context”

Abstract

A problem faced by some Logistic Support Organisations (LSOs) is that of forecasting the demand for spare parts, corresponding to equipment failures within the system. Here we are particularly concerned with a final phase of operations and the opportunity to place only a single order to cover demand during this phase. The problem is further complicated when the service logistics context can change during this final phase, e.g. as the number of systems supported or the LSO’s resources change. Such a problem is typical of the final phase of many military operations.

The LSO operates the recovery and repair loop for the equipment in question. By developing a simulation of the LSO, we can generate synthetic operational data regarding equipment breakdowns, etc. We then split that data into a training set and a test set in order to compare several approaches to forecasting demand in

the final operational phase. We are particularly interested in the application of Bayesian network models for this type of forecasting since these offer a way of combining hard observational data with subjective expert opinion.

Different LSO configurations were simulated to create a test dataset and the simulation results were compared with the various forecasts. The BN that learned from training data performed best, followed by a hybrid BN design combining expert elicitation and machine learning, and then a logistic regression model. An expert-adjusted exponential smoothing model was the poorest performer and these differences were statistically significant. The paper concludes with a discussion of the results, some implications for practice and suggestions for future work.

Keywords: Bayesian Networks, failure rates, spare parts forecasting, changing demand context

1. Introduction

The management and forecasting of spare parts for repairable systems is a vital part of support operations. This is particularly true for military equipment. For example, Moon et al. (2012) examine the forecasting of spare parts demand in a naval setting. Dekker et al. (2013) also clearly stress the importance of good demand forecasts. The usual methods applied are variations of time-series (Petropoulos et al. 2014). However, as Dekker et al. (2013) discuss, there are cases where time-series cannot cope well. Firstly, many parts do not exhibit a constant failure rate. Secondly, the usage context is unlikely to stay the same throughout the life of a supported system. Usage rate changes not only due to changes in the workload but also because of how many systems share the workload. The number of systems sharing the workload changes due to purchases and retirements, and the length of time for which some systems are undergoing repairs. This is where availability affects consideration of future failures: if periods of downtime are comparable to the designed time between failures of important parts, then equipment downtime becomes a driving factor affecting the frequency of failures. Consequently, the effectiveness of the whole support system itself becomes an indirect but important contributor to the

experienced failure rates. Finally, time-series cannot cope well when such changing conditions are combined with time-limited operations such as Search and Rescue (SAR), Disaster Relief, etc. The change in the demand producing context and the need for a single period demand forecast calls for more research in approaches to forecasting which might be better suited to such problems. A similar call is made by Dekker et al. (2013), to develop a forecasting method that explicitly takes account of installed base information:

"One could say that installed base forecasting is a kind of causal forecasting, in the sense that the forecast is not only made on the historic demand data but also on data about installed base aspects that trigger demand." (Dekker et al. 2013 p36) According to their definition, installed base refers to "the whole set of systems/products for which an organisation provides after sales service". Relevant information related to this definition includes maintenance and spare parts needed to support the systems, the service network with repair and stock locations, the maintenance concept, the age and the condition of equipment (e.g. for UAVs, the number of flying hours / usage), the lead times for spare parts and other logistic delays.

Additional factors that can affect the installed base functions include the environmental conditions, the number of operating hours and users' interventions such as decisions to change the geographical distribution of the operational systems or the repair capabilities at certain nodes of the support network. This thinking was indirectly supported by the study of Sherbrooke (2000) on the effect of the number of sorties and of the flying hours on the prediction of aircraft spares demand in Operation Desert Shield/Desert Storm in Iraq (1993-1996). In his analysis of more than 700,000 sorties, Sherbrooke understood that he needed to control for factors such as material condition, aircrew proficiency and mission type.

In this paper, we investigate the final phase of operations of an LSO in which contextual factors, such as those mentioned above, can change, thus influencing failures and subsequent spare parts demand. This is an important problem in practice but one which has received little attention in the literature. A notable

recent exception in this regard is work by Rekik, Glock, and Syntetos (2017). While the focus of their work is on improving the level of adjustment made by the human expert, however, ours is on investigating the potential of an alternative approach, that of Bayesian Networks.

A useful review of spare parts forecasting was conducted by Boylan and Syntetos (2010). Within this, they suggested that the activities supported by a forecasting support system (FSS) (Fildes, Goodwin, and Lawrence 2006) could be split into three phases: pre-processing, processing and post-processing. These phases corresponded to problem classification, implementation of an appropriate forecasting approach and subsequent expert judgemental adjustment, respectively. They also noted that in practice, the use of both simple forecasts based on some kind of exponential smoothing and expert judgemental adjustment were widespread in spare parts forecasting. This helps to explain our inclusion of such an approach as a comparator to Bayesian networks.

The particular problem considered here can be categorised as a single-period, non-stationary forecasting problem since we have to forecast spare part demand for a limited time-period ahead, during which the operational context can be very different to that which has been recently experienced. The literature concerning non-stationary forecasting problems suggests increasing the available relevant dataset by gradually collecting demand data from the new period, and applying Bayesian (Popović 1987; Huang, Leng, and Parlar 2013) or time series (Alwan et al. 2016) updates to the first moment of the assumed distribution . However, such methods are not suitable for the problem considered here due to its single-period nature. For example, in an overseas military operation, where the lead times are quite long, only a single order can usually be made before any additional data can be collected, and therefore the ability to regularly update the forecast of remaining demand in the light of fresh demand information is of little value.

In order to provide comparisons with the forecasts developed using BNs, we have chosen logistic regression and a forecast employing expert adjustment away from a single exponential smoothing baseline. The logistic regression model can take account of the changing contextual factors and, like the BN models, estimate the

probability of an equipment failing during a time interval within the final period of interest which can then be scaled up to create a demand forecast. The expert-adjusted forecast relies on the expert's judgement to take suitable account of the information available regarding the contextual factors. Full information was made available to the experts concerning the values taken by the contextual factors during earlier operating periods, together with the associated baseline forecasts and realised demands. They were then presented with the values taken by the contextual factors corresponding to the final period along with the SES baseline forecast and asked to predict the demand. Such contextual information is sometimes described as 'market intelligence' in the context of sales. Our reason for including this comparison was motivated by our expectation of this being typical of current practice. As well as Boylan and Syntetos (2010), many other authors, including Franses and Legerstee (2010), Fildes et al. (2009) and Klassen and Flores (2001), make clear that many of the model-provided demand forecasts are often then adjusted by the decision makers/subject matter experts before arriving at the final figure to be used, "ostensibly to take account of exceptional circumstances expected over the planning horizon" (Fildes et al. 2009 p.3).

Our main interest in this paper is in exploring the application of BNs (Pearl, 1988) to this problem. These provide a powerful and flexible approach to reasoning under uncertainty. There have been a number of studies investigating the use of BNs in related fields including reliability (Langseth and Portinale 2007), maintenance (Weber, Jouffe, and Munteanu 2004; Weber and Jouffe 2006), system testing in manufacturing (Chan and McNaught 2008) and supplier selection (Hosseini and Barker 2016). However, we have not found any application to the kind of logistical support problems outlined here.

We present a comparison of results generated from BNs developed in different ways along with those generated from more traditional forecasts – a statistical regression model and expert predictions adjusted from a fixed exponential smoothing forecast. The comparison makes use of data from a simulated scenario of a logistics support network of a fleet of generic UAV systems. Differences arise due to the way in which the different methods make use of

available information on the demand and support defining context. Furthermore, as we discuss later, BNs have the potential to provide not only predictions of the failure rates, but also of other factors such as the time to repair and to resupply which are needed for Multi-Indenture Multi-Echelon (MIME) spares optimization models.

The rest of the paper is organised as follows. In Section 2 we describe the simulation that we built in order to generate the data needed to develop the demand prediction models that we compare and also for the evaluation of their performance. In Section 3 we describe the forecasting methods employed to predict the number of failures in the final phase of operations. Section 4 contains the results from the simulation runs and a comparison of the various models' forecasts. These are discussed before some final conclusions are drawn and potential future work outlined.

2. Simulated system

Given the lack of readily available data of the kind needed to develop and test our models, and the likely sensitivity of such data even if it were available, it was necessary to simulate a Logistics Support Organisation (LSO) instead. In this section we describe the nature of the LSO, the scenario chosen for investigation and the generation of data for model building and subsequent testing.

2.1 Simulation of the LSO

The simulation (see Figure 1) concerned the support provided to a small fleet of generic Unmanned Aerial Vehicles (UAVs) that are used for surveillance at a single Forward Base (FB). The Logistics Support Organisation (LSO) was composed of a Forward support level (FORWARD) at which broken down items (Line-Replaceable Units (LRUs)) that make the UAVs non-operational are replaced with new ones from the inventory, and a Central repair level (CENTRAL) at some distance from the FB where the inventory of spares is kept and repairs are performed on the broken down items (the LRUs). The scenario was intentionally kept simple, so only corrective maintenance has been considered. Again, for the sake of initial simplicity, the Equipment Breakdown Structure (EBS)

of a generic UAV unit was composed of only a single LRU that could be repaired at the CENTRAL depot by the replacement of a single Disposable Part (DP) kept in the same store as the LRUs. Furthermore, we did not consider the case where systems' innate failure rates change with age. Finally, even though in real-life situations the spares demand might be intermittent, in order to get enough data, we simulated a UAV system that has breakdowns each month.

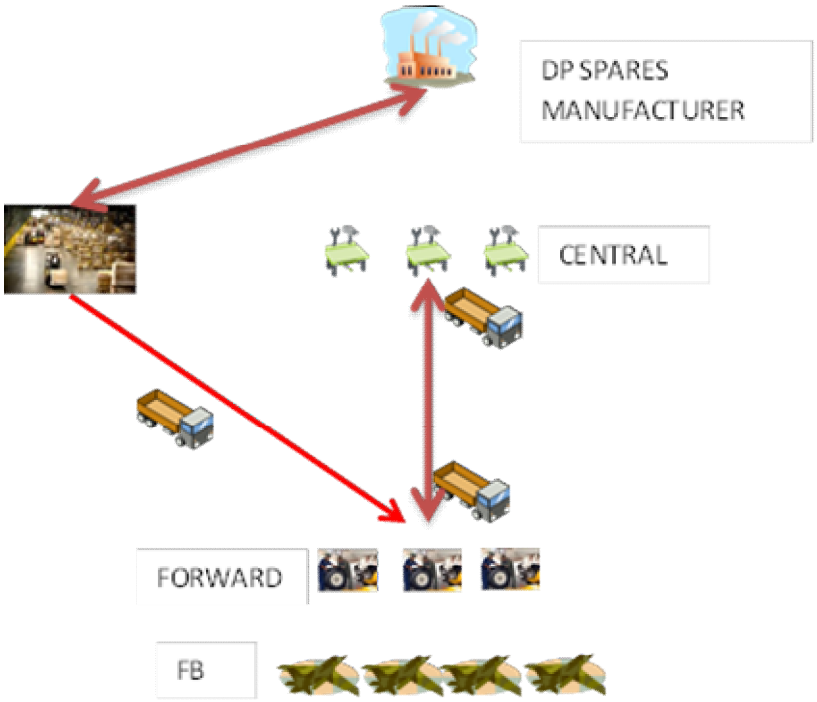


Figure 15: The simulated Logistics Support Organisation

The main objective of the LSO is to provide logistical support to a number of UAVs in their air-surveillance operations. In the assumed scenario, each UAV has a nominal Time on Task (TOT) of four hours, after which it has to land for a quick refuelling. If another UAV is available then it will take off; if not, the same UAV will be used again. The operational demand is to cover an area assigned for aerial surveillance by a single unit for a given proportion of the day, each day. For example, if the operational demand is to cover 4/5 of the day, since either there is no need to fly during night hours, or a different group takes over that period, then the operational demand (OpDem) is 4/5. Because of the importance of the air-surveillance function, there is always a mechanic assumed to be waiting to

help in case of a breakdown (B). If a breakdown occurs, another UAV takes off if one is available, and the grounded UAV is taken over by the mechanic who starts the diagnosis procedure. The duration of this procedure depends on the skill level of the mechanic, but we assume that the fault is always a single one and is always found correctly. After the diagnosis is over, an order for a spare is given at the CENTRAL depot. The spare takes some time to be located and acquired by a driver and is then brought to FORWARD. The mechanic replaces the faulty LRU with the spare, making the UAV available again. The LRU is then transported back to CENTRAL by the mechanic and the driver in order to be repaired. There are three available workbenches (W) at CENTRAL which are used for diagnosis and repair of the faulty items. The same mechanic is assumed to undertake the diagnosis and repair on one of the available workbenches and brings the LRU in a usable condition back to the LRU inventory, provided there is a DP in stock. Due to the assumed high cost of a DP, the depot uses an (S-1, S) inventory policy and thus initiates a resupply order whenever there is a single DP unit removed from the DP inventory.

2.2 Scenario for dataset generation

The chosen scenario involves a single iteration of the following consecutive eight phases (**Table 20**):

Table 20: Scenario Phases

Phase	Duration (Months)	xSLRU	xSDP	xNU	xNM	xNTr	OpDem
1	3	3	3	2	2	1	4/5 of a day
2	3	3	3	3	3	2	4/5 of a day
3	4	4	5	4	3	3	4/5 of a day
4	3	4	6	3	2	3	4/5 of a day
5	3	3	3	2	2	1	5/5 of a day
6	3	3	3	3	3	2	5/5 of a day
7	4	4	5	4	3	3	5/5 of a day
8	3	4	6	3	2	3	5/5 of a day

The assumed story behind the phases shown above is that during the 1st phase when operations started, there were two UAVs (xNU = 2) deployed with a mission to provide an air-surveillance function for the Operational Demand (OpDem) of

4/5 of a day. For the manning of the LSO in the 1st phase, there were two mechanics deployed (xNM = 2) and one driver (xNTr = 1), while the initial spares stock levels were three LRUs and three DPs (xSLRU = 3, xSDP = 3). The UAVs were flown by an equal number of operators with an initially sampled level of proficiency. As the operations built up in Phase 2, an additional UAV was deployed along with an additional driver to help with the transports of the spares and the mechanics. This situation lasted for three months and was followed by Phase 3, a four months phase when a 4th UAV was deployed along with an additional operator and driver. The spares holdings of LRUs and DPs were also increased at the beginning of Phase 3. In Phase 4, one UAV is withdrawn along with its operator and a mechanic. In Phase 5, the OpDem had to be increased to full 24hrs surveillance, although at the same time, one UAV was assumed to be failed beyond repair. In addition, it was assumed that one operator, two drivers and some spares were transferred out of the LSO. Further changes of this nature affecting the LSO's configuration were assumed for Phases 6 to 8, as shown in Table 1.

Records of take-offs and landings, of break-downs, of repair and re-order incidents, of on-hand (OH) and due-in (DI) spares and of number of deployed UAVs, mechanics and operators were kept from the single run of the consecutive eight phases, just like the records that would be kept in the relative logs of real operations. Furthermore, variables that can affect the incidents and the duration of diagnosis, repair and transport were also recorded. Such variables were the environmental conditions, the operators' skill levels/ experience, the mechanics' skill level / experience and their workload level.

2.3 Simulation of test data to allow forecast comparison

The end of Phase 8 provided the initial conditions for a follow-on ninth phase of six months' duration that was used to evaluate the performance of the demand prediction models. Our interest is in how well we can provide demand predictions when the failure-context factors are about to change. Consequently, Phase 9 could take different courses in order to represent a range of changes likely to be experienced in practice. Therefore, we simulated 18 different possible

configurations of Phase 9, none of which exactly replicate any of the earlier phases. These 18 configurations are listed in Table 2.

Table 21: The sample of LSO configurations that constituted the test dataset

xSLRU	xSDP	xNU	xNM	OpDem	Env
3	3	2	2	1	30%
3	3	3	3	1	50%
4	5	3	2	1	70%
8	8	3	2	1	50%
4	5	4	2	1	50%
3	3	4	2	2	30%
3	3	3	2	2	50%
8	8	4	2	1	30%
4	6	2	3	1	50%
3	3	4	2	2	70%
4	5	2	2	1	30%
4	6	4	3	2	70%
8	8	3	3	2	70%
4	6	3	3	2	50%
8	8	4	3	2	70%
4	5	4	2	2	50%
4	5	2	2	2	50%
4	5	3	2	2	30%

3. Forecasting Approaches Employed

Within the described LSO and operating context, there are many interacting factors to consider. This suggests the need for a modelling methodology that can take into account the effects of and the associations among the context defining variables. A natural modelling framework to consider here is that of Bayesian Networks (BNs). This is because within the problem being considered there are several random variables with probabilistic dependencies between them and BNs provide an efficient way of representing and manipulating such joint probability distributions. BNs also provide a flexible way of combining subjective expert opinion with observed data so that the same type of approach can be applied to situations with varying levels of available hard data.

The qualitative structure of a BN is represented by a directed acyclic graph (DAG), portraying probabilistic dependencies and independencies within the domain. This contains a great deal of information, even before we consider any probability distributions. The nodes correspond to variables of interest within the domain and arcs correspond to direct probabilistic dependencies. A fully specified BN, however, also requires a conditional probability table (CPT) for each node. These can be obtained from an appropriate dataset or elicited from a domain expert when insufficient data exists. Once complete, a BN offers efficient probabilistic inference over the domain of interest, allowing a decision maker to see how the probability distribution of some target variable is likely to change in response to new observations or other relevant information. In our specific case, our main value of interest is the probability of experiencing a failure incident (binomial variable “FRT” in Table 3) at any specific hour. Under the assumption of a Poisson process we get the required mean number of failures for the duration of the forecasting period by multiplying the acquired rate figure by the respective 4320 hours included in the 6 months of the final phase. We believe that the Poisson process is a valid assumption in these cases, given that we have also assumed that the operated systems do not degrade and that the only reason for the change in the failure rates is the context formulated by the support operations and the operational demand.

In order to provide a comparison with the BN predictions, we also provide forecasts using two other methods. The first is a logistic regression, which will also try to account for the relationships between the contextual factors and the observed number of failures. The appropriateness of this type of regression model stems from the underlying random process which involves the generation of failed equipment. The output, as for the BNs, is the probability of experiencing a failure incident in any specific hour.

The second type of additional forecast is the one most likely to be encountered in practice – human judgement. Since, along with the starting configuration for the ninth/final operational phase, our judges were also supplied with the simple exponential smoothing forecast available at the end of the eighth operational

phase, this could be described as an expert adjusted forecast, with adjustment being made away from the fixed SES forecast.

A BN can be developed in different ways, using different combinations of human expertise and data (Korb and Nicholson, 2004). When developed entirely from a dataset, it is said to have been learned from that dataset. This entails both the structure of the network, i.e. the DAG, and the associated CPTs being derived from the dataset. While obtaining CPTs from a dataset is relatively straightforward, deriving the structure is much more involved. This is primarily due to the huge number of DAGs which can be built from even a relatively small number of variables. Since there are also potentially a large number of DAGs which can represent the dependence structure of the joint probability distribution of interest, albeit some more efficiently than others, we need a way of identifying an efficient DAG for our purposes.

Instead of deriving a BN's structure from data, another common approach is to elicit the structure from a subject matter expert. In particular, making use of their causal knowledge of the domain, human experts can often quickly identify an efficient DAG. Such a DAG is usually easier to understand and so explain to decision makers. However, this DAG may omit subtle or less obvious relationships within the domain. In such a case, a BN learned from data might outperform the expert-elicited 'causal' BN.

A hybrid approach can also be adopted. Here, the subject matter expert (SME) can provide an initial DAG and some constraints on the structure which is then built upon by an automated machine learning algorithm. This ensures that key relationships are communicated in an understandable way and that more subtle effects are not missed.

As should now be clear from this discussion, different types of BNs can be applied depending on the quantity of data available. Of course, when datasets are plentiful, many approaches are possible, including, for example, artificial neural networks. The situation is very different, however, when data are sparse. Their ability to cover the spectrum of data availability is one of our key motivations for

employing BNs in this paper. They still allow a logical forecasting model to be developed for new products or situations with very limited historical data.

In order to develop forecasts using the approaches described above, we began by identifying candidate variables. Key to our thinking was to use the kind of data we could expect to be recorded in log-books across the LSO.

3.1 Grouping of the variables

The failure rate of repairable systems and the associated demand for spare parts is affected not only by how many systems we have deployed but also by their availability. This makes the factors that affect the systems' operational availability an important set of variables that indirectly contributes to the experienced number of failures.

Additionally, we can expect the failure rates of the systems to be affected by a number of factors such as the conditions in which each one works, the skill level of the operator, etc. Hence, we can identify three groups of "causal" variables. Each of these groups can be considered individually at each level of the LSO, including the level where the supported systems work. These groups are:

1. Factors related to the amount of use of the supported system – the "failure creators",e.g. the operational demand for number of missions in a given day, and the time required on task.

2. Factors that make the usage more prone to failure - the "failure enhancers",e.g the environmental conditions, the number of hours that the system has flown without maintenance, and the level of expertise of the system's user such as the pilot.

3. Factors that affect the repair loop – the "repair loop characteristics", such as the time to repair a fault and the level of on-hand spares.

Eventually, we included the following variables:

Table 22: Nomenclature

OpRT: Operational Incident at FB, with values "Take-off" and "No new take-off"
xNU: The number of UAV units deployed

OpDem: Operational demand, with values 4/5 and 5/5 of a day
TOT: Time on Task; the realized continuous but discretized time on task of the UAV that performs the flight
PExp: The skill level of the operator (pilot) with three discrete values
Env: The environmental conditions with two discrete values, “OK” and “Not OK”
FRT: Failure Incident at FORWARD, with values “New Failure” and “No-New Failure”
Rdu: The duration of repair at FORWARD (discretized)
FlHbd: The number of flying hours since the last repair (discretized)
xNM: The number of mechanics deployed
MExpB: The skill level of the mechanic that took over the repair at FORWARD
QM: The percentage of mechanics that are idle
BWkld: The percentage of the FORWARD repair facilities that are occupied
xNTr: The number of drivers that have been deployed to do the transport from CENTRAL to FORWARD and back
QAdm: The percentage of drivers that are idle
WFRT: Workbench LRU failure Incident at CENTRAL, with values “New Failure” and “No New failure”
WRdu: The duration of repair at CENTRAL (discretized)
MExpW: The skill level of the mechanic that took over the repair at CENRTAL
WWkld: The percentage of the CENTRAL repair facilities that are occupied
ORT: Order for a resupply Incident, with values “New Order placed” and “No New Order placed”
Odu: The duration to be realised of the resupply that was ordered (discretized)
xSLRU: The nominal level of LRUs in the inventory
OhLRU: The on-hand level of LRUs
xSDP: The nominal level of DPs in the inventory
OhDP: The on-hand level of DPs
DiDP: The number of DPs which are on order but have not arrived yet (Due-in)

The variables in Table 3 that are highlighted in bold relate to incidents at the LSO levels in which the UAVs are used and supported. The other variables correspond to the three groups of contextual factors discussed earlier.

3.2 Expert-elicited BN

A BN of the problem situation was developed by first eliciting a DAG from a domain expert. This DAG displays the relationships believed by the expert to exist in the system. Such a human-elicited DAG can often be portrayed as a causal model since humans think naturally about relationships in a causal manner and this is in fact how we usually encourage experts to think when eliciting a BN DAG from them. Naturally, this predominantly causal form makes the model easier to

understand and explain to others. The DAG elicited from our domain expert is presented in Figure 2.

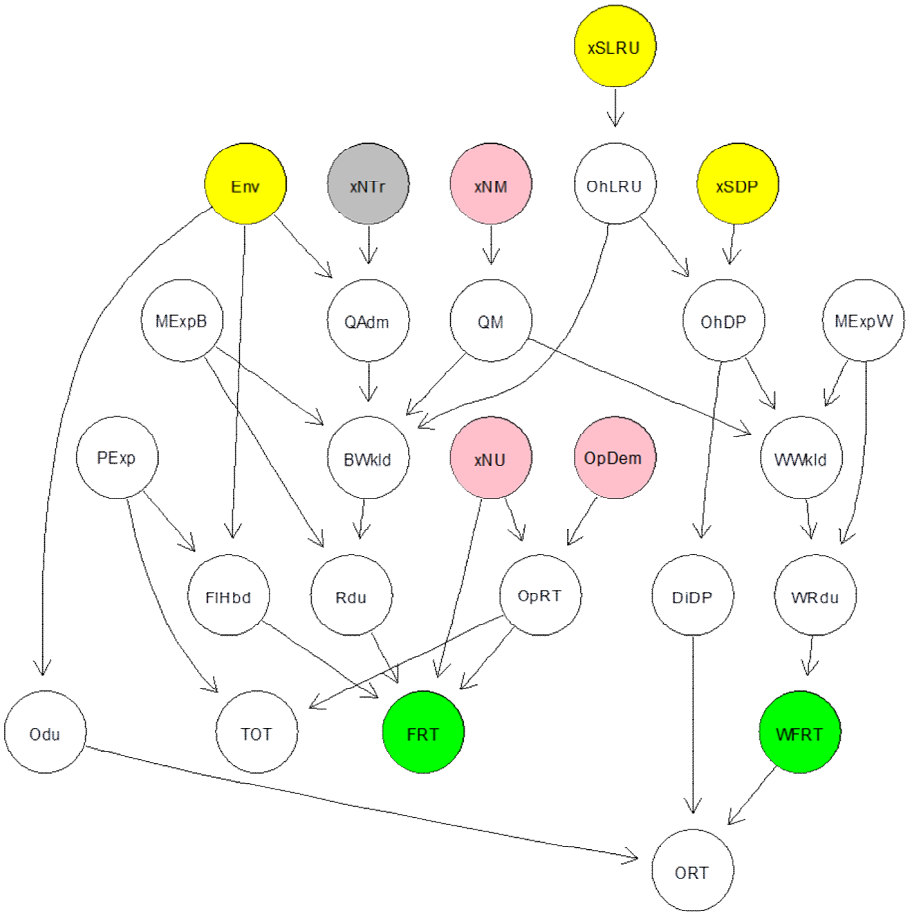


Figure 16: DAG of a BN model elicited from a domain expert

3.3 BN learned from data

It is important to realise that a BN learned from a dataset will not necessarily produce the same DAG as a BN developed using expert elicitation. The simulated log-book records can be used to obtain values for all the variables. Using the BN learning package in R called "bnlearn"[53] this sampled dataset of records from Phases 1 to 8 was fed into a score-based unsupervised learning algorithm. This applied the tabu search algorithm to 300 bootstraps and developed 300 networks that were averaged to form the final network. The scoring method employed the

[53] Developed and maintained by Dr Marco Scutari

Modified Bayesian Dirichlet equivalent uniform (MBDeu) score (Cooper and Yoo, 1999; Heckerman, Geiger and Chickering, 1995)

The above procedure produced the network displayed in Figure 3. The resulting graph is a representation of the joint probability distribution of the modelled variables.

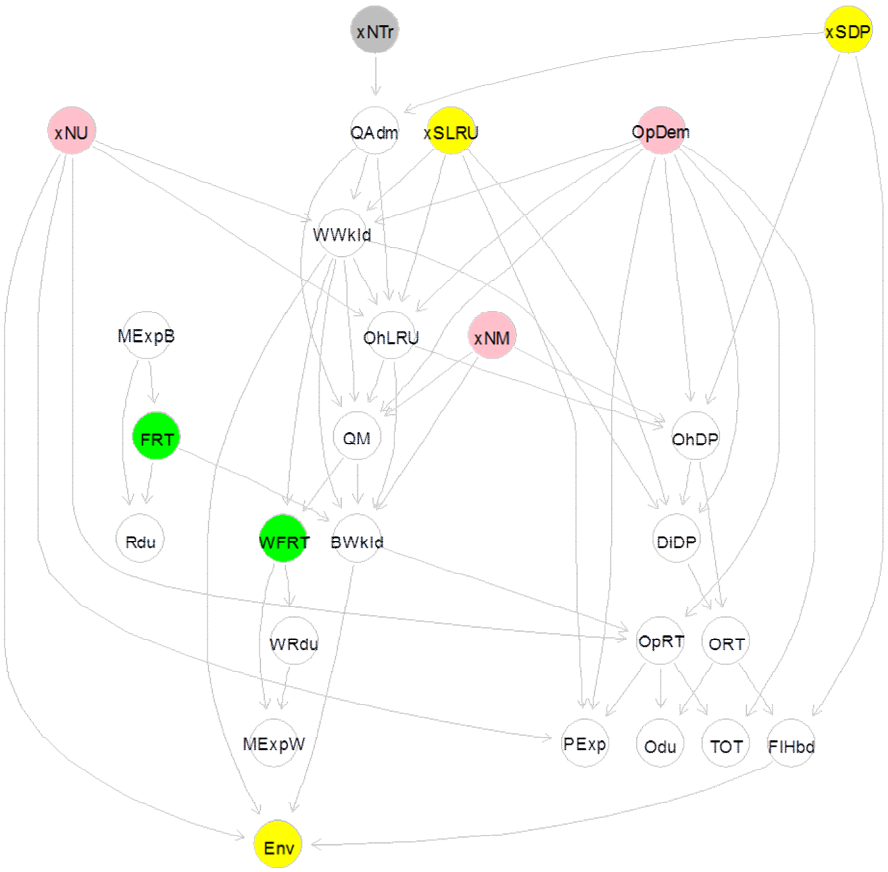


Figure 17: DAG of the BN model that was learned from the simulation training dataset

Note that the resulting model is not a causal BN since the causality assumptions are not met (see eg Pearl (1988)). However, it does provide an interpretation of the relationships / associations among the variables. For example the arc which connects xNU directly to OpRT and the arcs that connect the latter to the TOT indicate that the number of units operated (xNU) has a direct effect on the Operational Rate (OpRT), i.e. how often missing take-offs affect directly the resulting duration of any single take-off (TOT). Furthermore, most of the arcs are

directed towards the variables OhLRU (the on-hand LRU), WWkld (how busy are the repair workshops at the CENTRAL level) and BWkld (how busy are the workshops at the FORWARD level). This indicates that these facilities are key to the whole system.

3.4 Hybrid BN

A hybrid BN was developed in order to try and obtain the best of both worlds. Ideally, we would like to have the understandable nature of the expert-elicited BN combined with the ability to learn less obvious relationships provided by the learned BN. To develop this hybrid, we began with a simplified version of the expert-elicited BN and used this as a starting point for the machine learning algorithm which was employed to develop the learned BN. This constrains the final DAG to incorporate the expert-elicited components but allows additional relationships to be included alongside that.

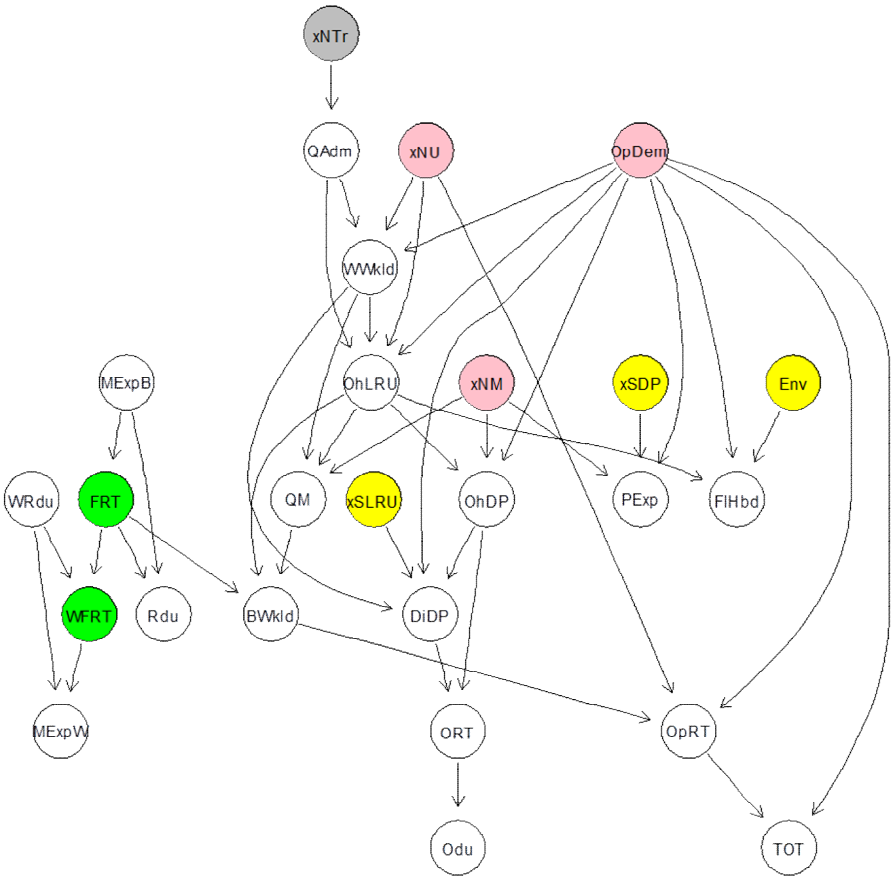


Figure 18: DAG of a hybrid BN, combining expert elicitation and machine learning

As should now be evident, in order to obtain the joint probability distribution of the variables chosen to model the system, many different factorizations are possible, corresponding to different DAGs. However, some of these are simpler and more efficient, depending on the actual relationships between the variables. For each of these DAGs, the simulated data were then used to calculate the Conditional Probability Tables (CPTs) associated with them (Korb and Nicholson, 2004).

3.5 Logistic regression model

The logistic regression model derived from the first eight phases of the simulation training dataset was the following:

$$logit(FRT) = b_0 + b_1 OpDem + b_2 EnvCond,$$

where FRT corresponds to the occurrence of an equipment failure, OpDem represents the level of operational demand (in this scenario, how much of the day an equipment is required for) and EnvCond represents the severity of environmental conditions.

The coefficients of b_0, b_1 and b_2 are -4.5273, 0.4418 and 0.1836, respectively, where the reference settings of the variables are '4/5 of a day' for OpDem and 'OK' for EnvCond. In order to forecast demand for Phase 9, where the state of the EnvCond variable is not yet known but we have a probability distribution for it, the forecast uses a weighted average of the output obtained with the two possible values of this variable.

3.6 Expert-elicited forecast

In order to construct this forecast, four domain experts were consulted. Each was talked through the scenario implemented in the simulation and provided with the same information. This consisted of the configurations of the eight initial phases of operation and the resulting number of failures observed. Each was then asked to provide a forecast of the number of failures expected for a final ninth phase of operations given the LSO configuration and the simple exponential smoothing

estimate, purely based on the previous eight phases and independent of the Phase 9 configuration. The fixed SES forecast was obtained using the "tsintermittent" R-package and provided monthly predictions with a smoothing factor of 0.2. 18 different possible configurations were considered for Phase 9 and each expert provided an individual forecasts for each of these. The mean of the four forecasts was then taken to represent the expert-elicited forecast for each Phase 9 configuration.

4. Results and Discussion

4.1 Results from the simulation and the forecasts

Results from the various forecasts are shown over Figures 5 and 6 in order to reduce the amount of cluttering in the overlaid plots. In each figure, the same set of 18 boxplots are reproduced to show the distribution of the Phase 9 failure rates across 100 simulation replications for each of the 18 configurations. The boxes in each case include the inter-quartile range of the number of failures from the 100 replications. The crosses indicate outlying values in the simulation results. Overlaid on each boxplot are the forecasts for that Phase 9 configuration. In Figure 5, forecasts from each of the three BN models are displayed in addition to the boxplots of the simulation results. In Figure 6, the logistic regression and expert-adjusted forecasts are given in addition to the simulation boxplots. The vertical axes of these figures record the number of failures for Phase9, either observed from the Phase 9 simulation results or forecast by one of the considered models. The 18 Phase 9 configurations are arranged in increasing order of the median number of failures obtained from the 100 replications of each of them.

Apart from the indicative differences evident within Figures 5 and 6, we tested for significant differences in the forecast accuracy, as measured by the Absolute Relative Error (ARE) score:

$$ARE = \frac{|Y - Y'|}{Y},$$

$(Y: Actual\ number\ of\ failures,\ Y': Estimated\ number\ of\ failures)$

The AREs of the various models were compared using the Friedman non-parametric test over the 18 configurations of simulated futures, each such configuration being replicated 100 times. Friedman's test was chosen instead of its parametric equivalent, ANOVA, since we cannot assume sphericity in the measured absolute relative errors (Demšar, 2006). The test's p-value was less than 1%, providing evidence to reject the null hypothesis of no difference in the forecast accuracy between methods at that significance level. Furthermore, we applied a post-hoc Nemeneyi test to rank the models (Garcia and Herrera, 2008). This test showed that the order for the accuracy performance of the examined models (from best to worst) was the unsupervised BN learned from data, the hybrid BN, the logistic regression model, the causal BN with its DAG elicited from an SME and the SME adjusted SES, with a critical distance between ranks of 2.098 at the 1% significance level and mean ranks of 178.7, 211.2, 249.6, 254.1 and 359.1, respectively, i.e. the accuracy performance of all forecast methods are significantly different at the 1% level.

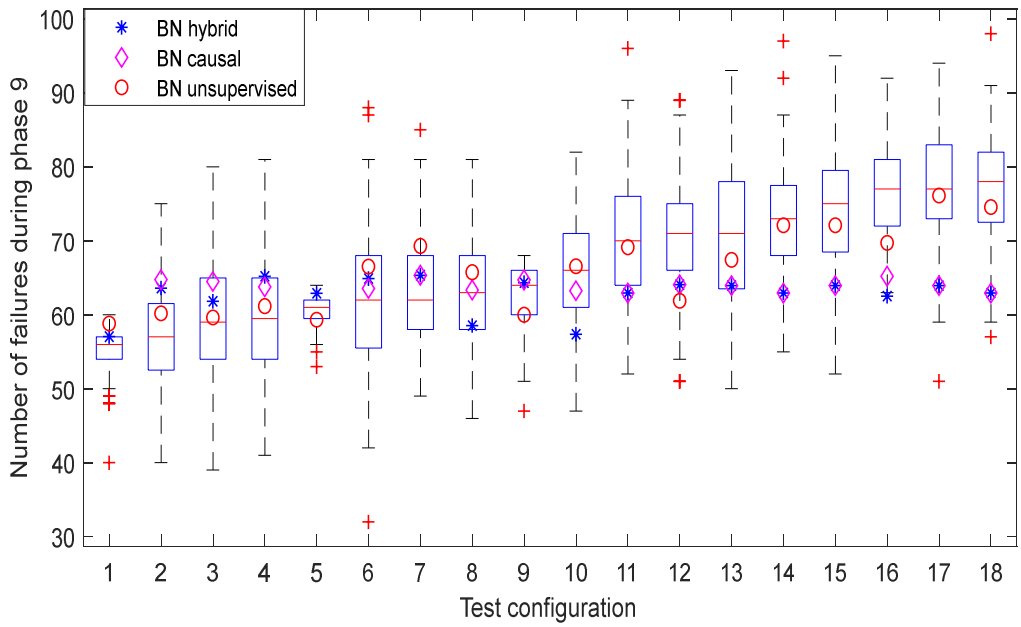


Figure 19: A comparison of the BN models' forecasts and the simulation results

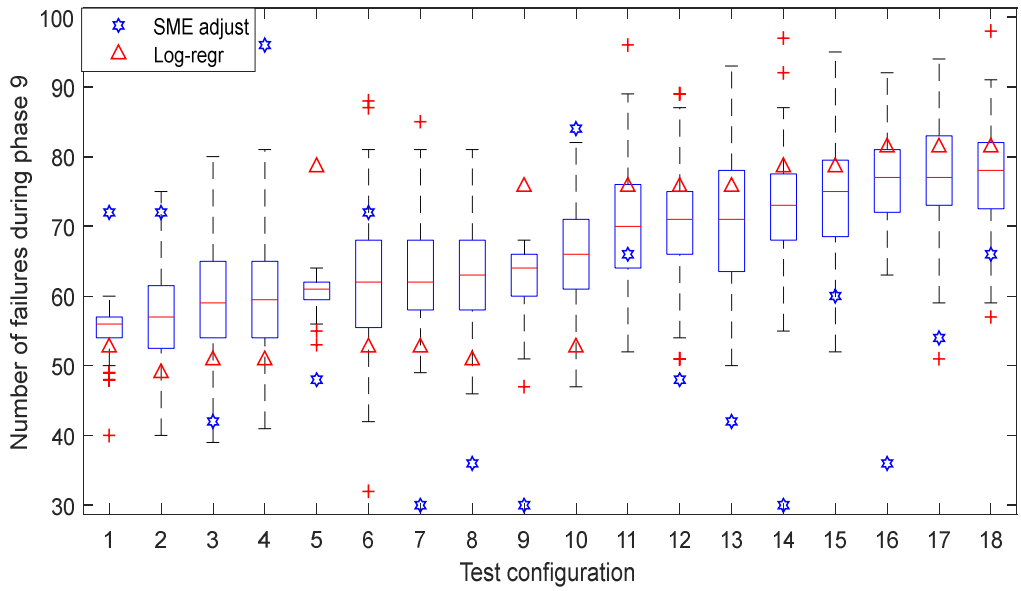


Figure 20: A comparison of the regression and the mean SME forecasts and the simulation results

4.2 Discussion

From Figures 5 and 6, and the subsequent statistical analysis, we can see that the Bayesian network models outperformed both the expert-adjusted forecast and the logistic regression model. Furthermore, of the three approaches to BN construction considered, the BN developed by machine learning algorithm performed best, followed by the hybrid BN and then the expert-elicited BN. Of course, we need to speculate on why the BN models did not perform even better.

Predicting failures with the BN and logistic regression models essentially treats the situation like a classification problem, taking some characteristics of the period during which a failure occurred in the training data and using these to help estimate the probability of a failure when such characteristics are present at the start of a new period in the test data. However, there could still be differences in a time period's initial conditions outwith these characteristics, having some influence on demand. Simple aleatory or random variation of the Poisson failure process is also going to play a part.

Regarding the dataset, one of the decisions that needed to be made was on the time periods that would be used in the collection of the data and in the subsequent development of the regression and the BN models. A useful framework to consider in this regard is the Aggregate-Disaggregate Intermittent Demand Approach (ADIDA) (Nikolopoulos et al (2011)). The method mainly addresses the problem that models have when there are intermittent demand time series.

Fildes et al (2009) note that provision of a statistical forecast to the expert is likely to influence their thinking which may result in under-adjustment from that forecast, based on Tversky and Kahneman's (1974) anchoring heuristic. That could have been true in our experiment as we provided our experts with the SES forecast. However, since this forecast was fixed and known to take no account of planned changes to the LSO configuration, it is likely to have had a weaker effect than a forecast which did account for planned changes. In fact, looking at the mean performance of the experts in Figure 6, the magnitude of adjustment does not appear particularly small but the direction of the adjustment is often wrong. This appears to echo Sterman's observations on the difficulty of incorporating feedback into our thinking. The nature of the repair system considered gives rise to dynamic feedback effects which can sometimes create counter-intuitive behaviour and present difficulties for human judgment (Sterman, 2000).

The benefits of using a BN to forecast the number of failures are not limited to that immediate forecast. Other variables can also be queried which is useful in itself and also for providing explanations. In Phase 9 of the simulation, for example, we found that if there are 4 UAVs deployed for an operational demand of 24/7 surveillance, which are supported by 3 mechanics and by an investment on 3 DPs, a TOT of at least 3 hours has a probability of 0.85 while the probability of such a desired event increases to 0.92 if one more mechanic is deployed and the level of DPs is increased by 2. Furthermore, a TOT of at least 3 hours has a probability of 0.91 when there are 3 mechanics and 5 DPs but with one less UAV, i.e. 3 instead of 4. As another example, our BN suggested that the duration from the time that a DP resupply order was placed until it arrived was most probably less than 210 hours, while the median value experienced throughout Phases 1 to

8 was 215.7 hours. This is useful since MIME optimization models make use of time durations, like time to repair, time to transport / resupply, etc., which are used in order to calculate the parameters for the pipeline levels' probability distributions. A final example for the intuition that the development of the BN can offer is related to a logical fallacy that decision makers tend to make due to the human limitations in seeing the support system as a whole. We have experienced cases in which the decision makers, in order to maintain the required fleet availability in the face of anticipated increases in operational demand, they suggest the deployment of more units. In our case, Phase 8 ended with an operational demand for a unit to be in the air 24/7 and 4 UAVs deployed. In the following table we see what we should expect if during phase 9 the decision makers deploy 2 UAVs and what if instead they deploy 4 UAVs without though affecting any parameters of the repair or the resupply configuration of the support system. In the table's first column (**Table 23**) we have these two questions which we examine under three different possible environmental conditions (30%, 50% and 70% of Phase 9's 6-months environmental conditions to be ok), while on the fourth column we have the percentage of the day that the decision makers should expect to actually have a UAV in the air. What we observe is that by operating 4 UAVs (3rd column rows 4 to 6) the percentage of time we actually have one in the air is less than when 2 UAVs are deployed (rows 1 to 3). The cause can be inferred from the two last columns. When deploying 4 UAVs without sufficiently amending the repair and resupply configuration of the support chain, the jobs both forward and at the repair shop increase to a level such that the actual flights performed are reduced.

Table 23: Additional BN queries

		OpRT	**BWkld**	**WWkld**
Phase 9 - alternatives	**Env OK**	*Flying*	*Working*	*Working*
OpDem:2 - U:2 - M:3	*30%*	97.82%	36.45%	60.79%
	50%	97.71%	38.13%	61.27%
	70%	97.58%	40.09%	61.83%
OpDem:2 - U:4 - M:3	*30%*	92.59%	75.37%	78.26%
	50%	93.12%	75.54%	78.15%

	70%	93.74%	75.75%	78.02%

Naturally, using simulation data can be criticised as being less realistic than using real-life data collected from an LSO. The acquisition of real-life data would require access to multiple logbooks from the different nodes within the LSO and subsequent cleansing and synchronising of that data which would nearly always be of a sensitive nature. The main advantage of using real data in studies such as this one would be the increased credibility of the results, particularly in the eyes of practitioners. However, for the purposes of comparing forecasting approaches, the use of simulation offers real benefits. Since real data can be contaminated with all kinds of errors and contain anomalies which are unrepresentative, the use of simulation provides a control to remove such undesirable effects. Reducing the level of noise in the data makes forecast comparisons more accurate and it is this comparison which is our primary interest. Furthermore, whereas the use of real life data would restrict us to just one realised future configuration of the LSO to make a prediction for, with simulation we can create many such possible future configurations. This provides a wider range of situations to compare the forecasting approaches over and increases the power of statistical testing when looking for significant differences between them. Finally, although the development of a simulation is not a trivial task, it may well still be quicker than the time that would be needed to collect and process the necessary real-life data.

However, we also need to reflect on the cleaner nature of simulation data when drawing any conclusions about the likely benefits arising from the use of any of the forecasting approaches in practice. The introduction of messier, real data is undoubtedly likely to cause the level of improvement obtained from using any of these approaches to be less than that indicated when using simulated data.

4.3 Implications for Practice

In a review of forecasting within supply chains, Syntetos et al (2016) note that many important problems faced by practitioners have not been addressed by academic research. We believe that the problem addressed in this paper comes close to falling in that variety. While there is little published work in this area, a

notable recent exception is (Rekik, Glock and Syntetos, 2017) which investigates expert judgmental adjustments from a statistical forecast in a finite-time horizon setting and proposes an analytical model to support this.

We believe that our initial investigation is useful to practitioners in that it shows that relying purely on human judgmental adjustments in such situations is sub-optimal and can be improved upon to some extent by an alternative approach. Our work suggests that approaches based on Bayesian Networks and machine learning are worth further investigation in problematic areas where the assumptions of traditional forecasting methods such as those based on time series analysis could be questioned.

As alluded to in 4.2, practitioners can often obtain additional benefits from the development of a BN to forecast a particular variable since it is a more general and flexible type of model. For example, military commanders might be interested in the probabilities of the Time on Task (TOT) duration of a typical mission under certain support settings (which can be entered as “evidence” in the BN model already developed). This helps to illustrate a useful advantage of BNs in this kind of setting – having developed a joint probability distribution across a set of variables, we can quickly use it to make inferences about variables other than the immediate forecast variable.

Several authors have established that human judgmental adjustments applied to statistical demand forecasts are common in industry (e.g. Klassen and Flores, 2001). Various cognitive biases, such as optimism bias, have also been postulated as influencing those adjustments (Fildes et al, 2009). However, most of this research has been conducted in the context of sales, where higher demand is generally desirable. When the context is instead demand for spare parts following equipment failures within the same organization, lower demand is desirable. This different framing of the problem may lead to different biases being at work or to different effects arising from the same biases. Practitioners should be aware of the need to take such framing effects into account.

5. Conclusions

In this paper, we have applied a novel approach to a problem which despite being of real practical relevance has received relatively little attention in the literature. The problem setting considered is that of an LSO, where an accurate forecast of spare parts demand is required, corresponding to equipment breakdowns within the system. However, the distribution of demand is non-stationary due to several contextual factors which can take different values in each time period. Furthermore, we are particularly concerned with the final phase of operations and the placement of a single order to cover demand during this single period.

In current practice, the most common approach to such a problem is that of unaided expert judgement or else expert judgment applied to adjust a relatively simple statistically based forecast such as single exponential smoothing. Our results showed the relatively poor performance of expert adjusted forecasts away from a SES forecast. Supplied with information regarding configuration changes to the LSO, forecast adjustments were often made in the wrong direction, possibly indicating counter-intuitive behaviour.

The BN-based approaches that we investigated, and particularly the machine learning BN, outperformed both the expert-adjusted forecasts and the logistic regression model. However, although the differences in performance were statistically significant, the level of improvement was less than we had anticipated. This might be due to both the presence of simple random variation from the failure generating process and the inherent dynamic feedback within the simulated system which poses a challenge to all of the approaches considered.

Boylan and Syntetos (2010) have discussed how it may be beneficial to adopt a Forecasting Support System for spare parts forecasting. We agree with them but suggest that the scope of such a system should be expanded to include and cater for a wider range of circumstances than those they discussed. The criteria considered during their initial pre-processing or classification phase, should be expanded to cover these new situations; e.g. the number of periods to be forecast, the presence and extent of contextual factors affecting demand, and the extent of market (or equivalent) intelligence available regarding the values of these factors. Such an expansion would also cater for the kind of problem

described by Dekker et al. (2013) and outlined in Section 1. Similarly, the range of approaches which can be used in the second processing phase needs to be expanded to suit the wider range of problems.

Finally, regarding future steps:

- Our simulation settings created failure data which were not intermittent. These demand data were sufficient to learn a BN to adequately model the examined variables. In future work, we will consider scenarios with intermittent failures
- We further need to investigate how frequently such a BN should be updated to take account of fresh data.
- We also plan to investigate the applicability of neural network approaches for this type of problem since neural networks lend themselves to problems where non-linearities are prevalent. However, it is not yet clear whether the kind of simulation data we have employed in this paper would be sufficient to train such a model adequately.

More realistic support problems will be investigated by increasing the complexity of the Equipment Breakdown Structure of the generic UAV and in that way we will also be able to use service level metrics in our evaluation criteria.