International Journal of Production Economics, Volume 209, March 2019, pp. 325-333
DOI:10.1016/j.ijpe.2018.06.017

# Using Bayesian Networks to Forecast Spares Demand from Equipment Failures in a Changing Service Logistics Context

Petros Boutselis and Ken McNaught

## 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.

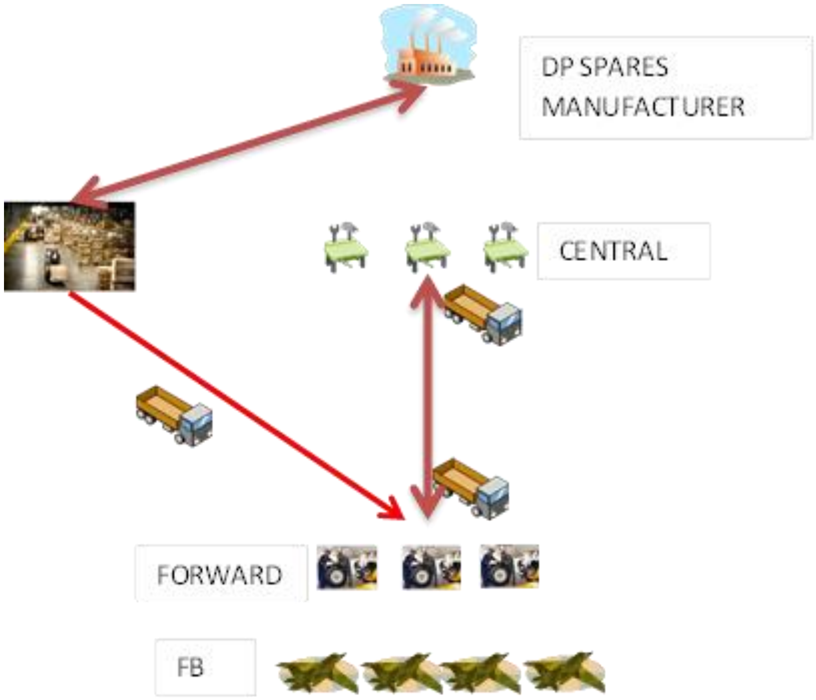

Figure 1: The simulated Logistics Support Organisation

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 1):

| 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 |

Table 1: Scenario Phases

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.

| 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% |

Table 2: The sample of LSO configurations that constituted the test dataset

## 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:

| |
|---|
| 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) |

Table 3: Nomenclature

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.

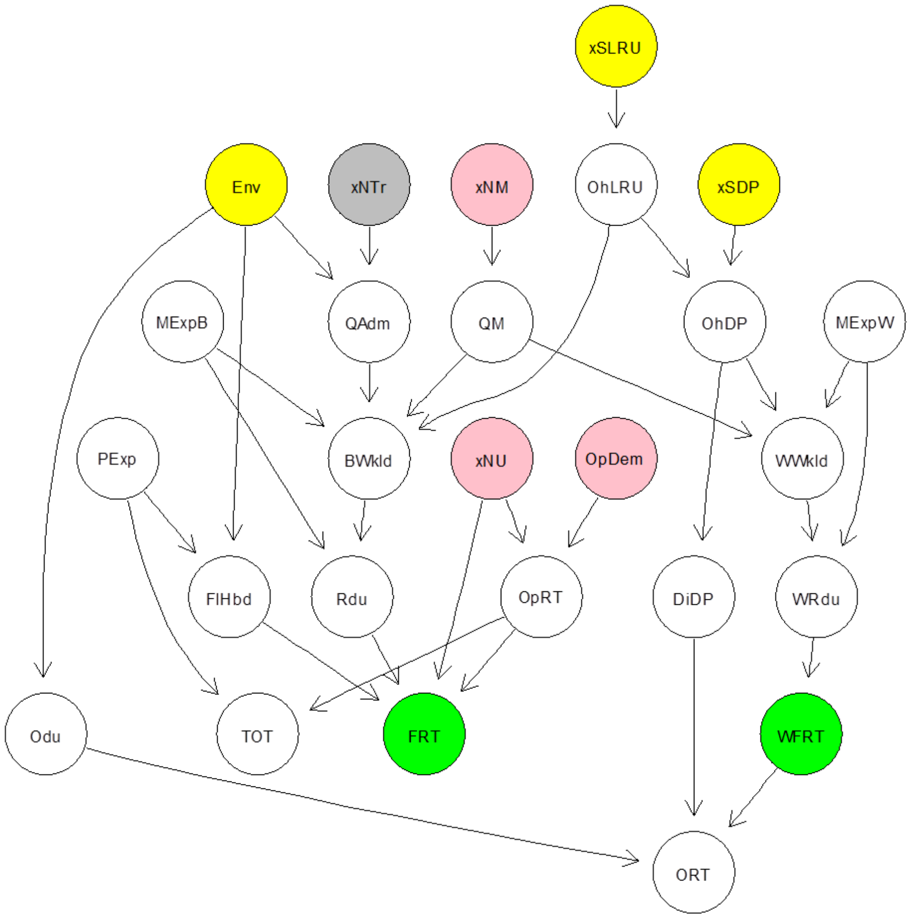

Figure 2: DAG of a BN model elicited from a domain expert

### 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”[1] 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 Modified Bayesian Dirichlet equivalent uniform (MBDeu) score (Heckerman, Geiger, and Chickering 1995; Cooper and Yoo 1999)

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.

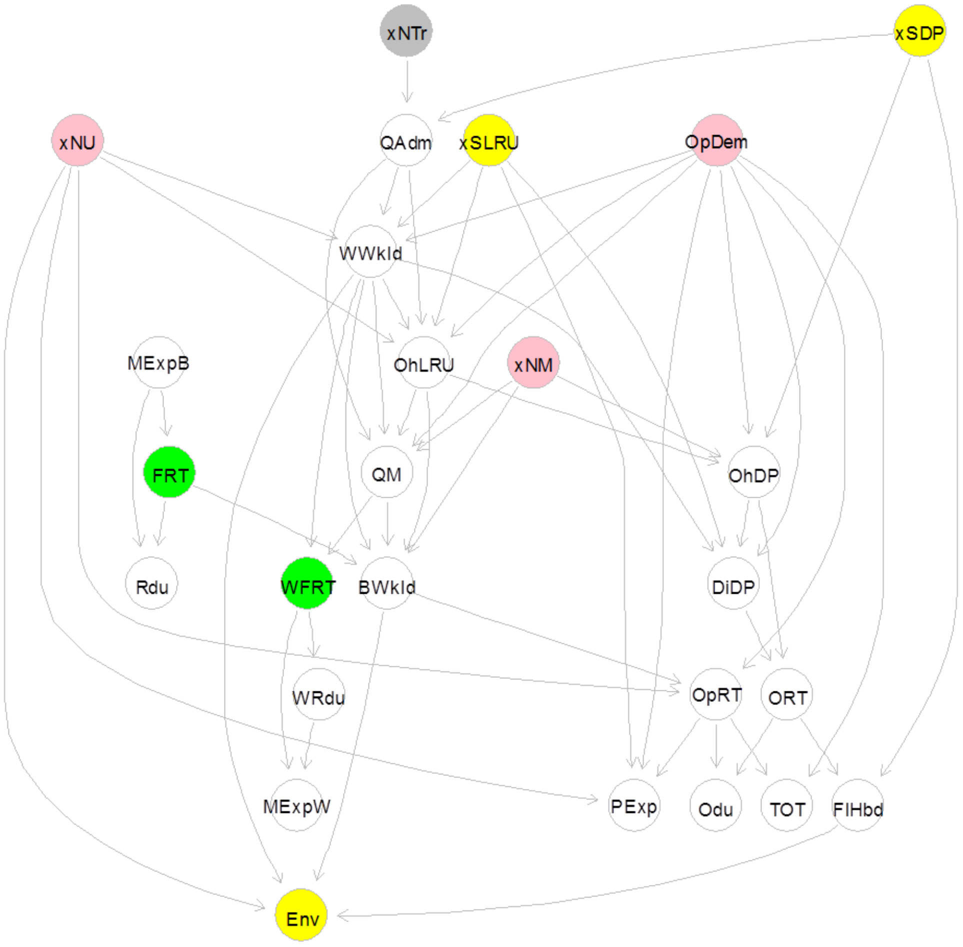

Figure 3: DAG of the BN model that was learned from the simulation training dataset

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

[1] Developed and maintained by Dr Marco Scutari

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.

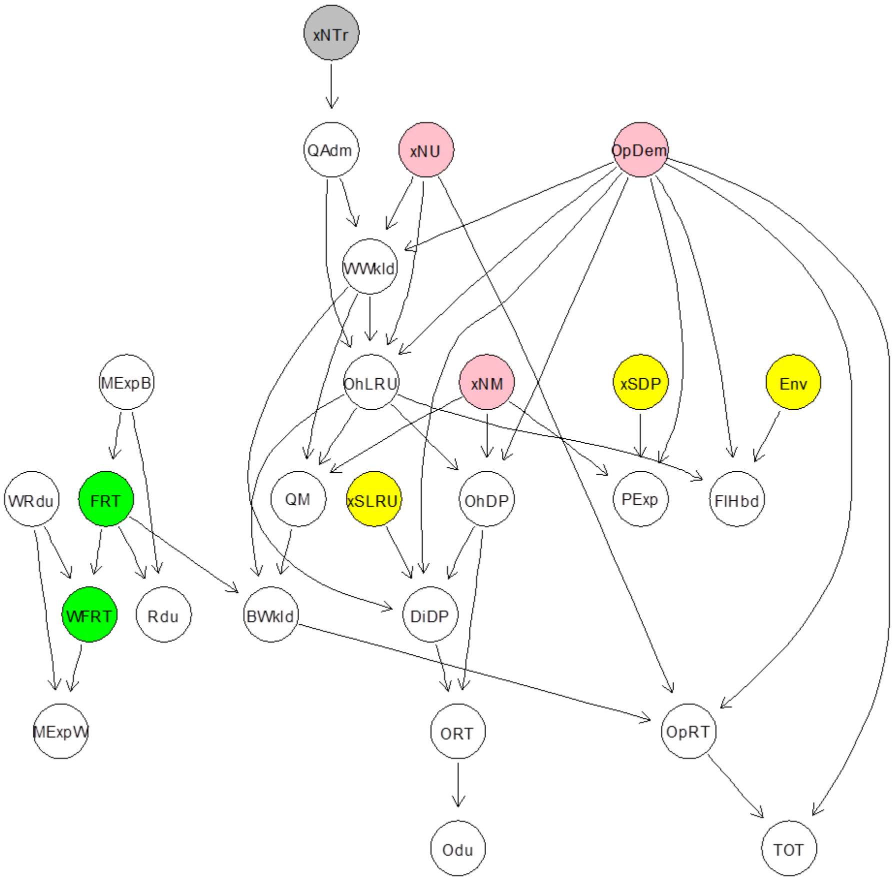

Figure 4: DAG of a hybrid BN, combining expert elicitation and machine learning

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 forecast~~s~~ 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:

$$ARE = \frac{|Y - Y'|}{Y},$$

$(Y\text{: }Actual\ number\ of\ failures,\ Y'\text{: }Estimated\ number\ of\ failures)$

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.

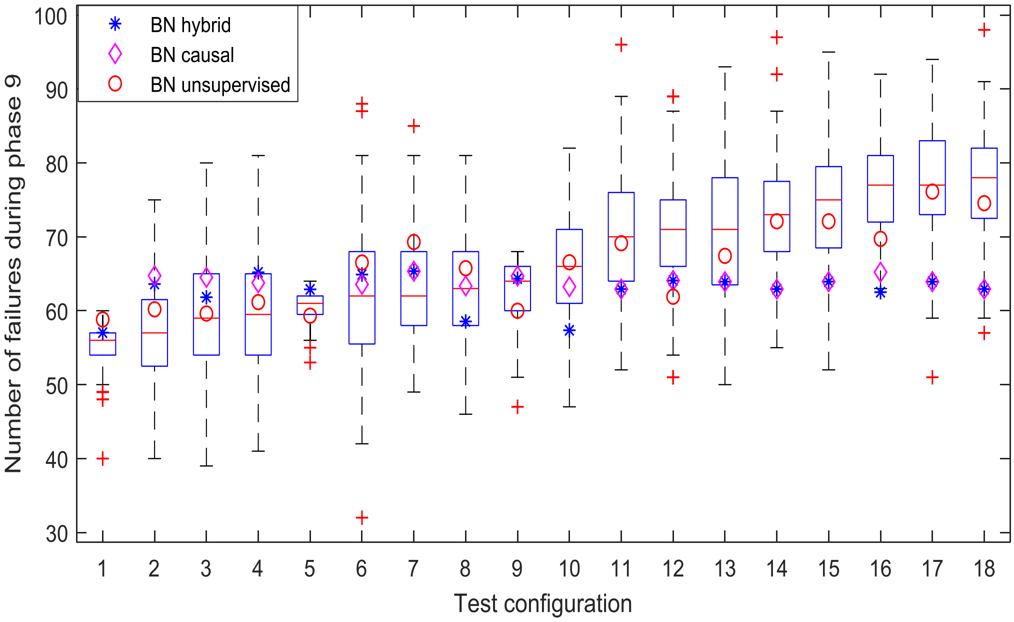

Figure 5: A comparison of the BN models' forecasts and the simulation results

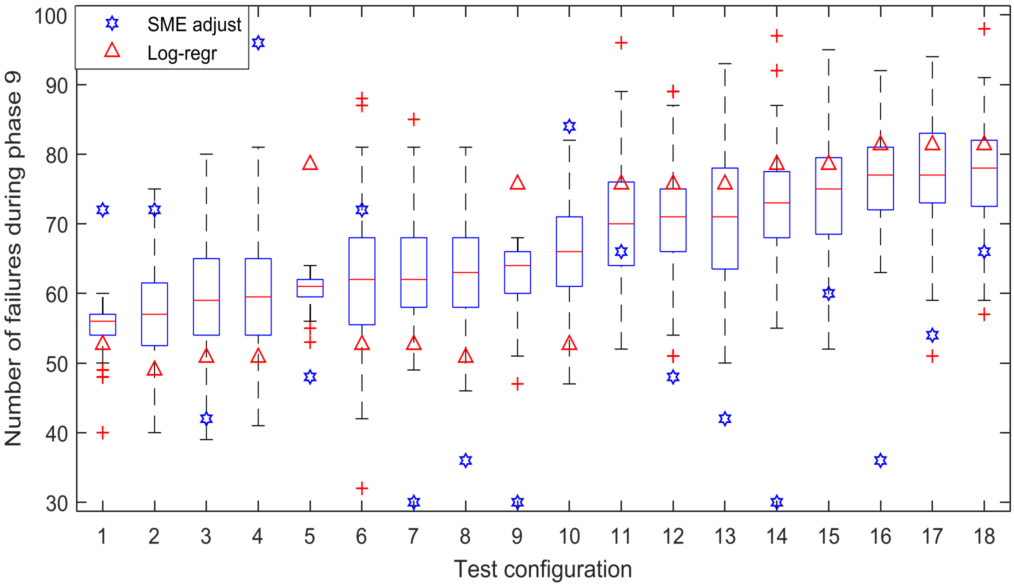

Figure 6: A comparison of the regression and the mean SME forecasts and the simulation results

### 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 4) 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.

| | | **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% |

Table 4: Additional BN queries

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.

**Acknowledgements**

We would like to thank the two anonymous reviewers and the Associate Editor for their helpful comments and suggestions on earlier versions of this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## References

Alwan, Layth C., Minghui Xu, Dong Qing Yao, and Xiaohang Yue. 2016. “The Dynamic Newsvendor Model with Correlated Demand.” *Decision Sciences* 47 (1): 11–30. https://doi.org/10.1111/deci.12171.

Boylan, John E., and Aris A. Syntetos. 2010. "Spare Parts Management: A Review of Forecasting Research and Extensions." IMA Journal of Management Mathematics 21(3): 227-237.

Chan, Andy, and Ken R. McNaught. 2008. Using Bayesian networks to improve fault diagnosis during manufacturing tests of mobile telephone infrastructure. Journal of the Operational Research Society 59: 423-430.

Cooper, Gregory F., and C Yoo. 1999. “Causal Discovery from a Mixture of Experimental and Observational Data.” *Proc. Fifthteenth Conference on Uncertainty in Artificial Intelligence {(UAI’99)}*, 116–25. https://doi.org/citeulike-article-id:3839452.

Dekker, Rommert, Çerağ Pinçe, Rob Zuidwijk, and Muhammad Naiman Jalil. 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): 536–45. https://doi.org/10.1016/j.ijpe.2011.11.025.

Demšar, Janez. 2006. “Statistical Comparisons of Classifiers over Multiple Data Sets.” *Journal of Machine Learning Research* 7: 1–30. https://doi.org/10.1016/j.jecp.2010.03.005.

Fildes, Robert, Paul Goodwin, and Michael Lawrence. 2006. “The Design Features of Forecasting Support Systems and Their Effectiveness.” *Decision Support Systems* 42 (1): 351–61. https://doi.org/10.1016/j.dss.2005.01.003.

Fildes, Robert, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. 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.: 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010.

Franses, Philip H., and Rianne Legerstee. 2010. “Do Experts’ Adjustments on Model-Based SKU-Level Forecasts Improve Forecast Quality?” *Journal of Forecasting* 29 (3): 331–40. https://doi.org/10.1002/for.1129.

Heckerman, D, D Geiger, and D Chickering. 1995. “Learning Bayesian Networks: The Combination of Knowledge and Statistical Data.” *Machine Learning, 20, 197-243* 243: 197–243.

Huang, J, M Leng, and M Parlar. 2013. “Demand Functions in Decision Modeling: A Comprehensive Survey and Research Directions.” *Decision Sciences*. http://onlinelibrary.wiley.com/doi/10.1111/deci.12021/full.

Klassen, Robert D., and Benito E. Flores. 2001. “Forecasting Practices of Canadian Firms: Survey Results and Comparisons.” *International Journal of Production Economics* 70 (2): 163–74. https://doi.org/10.1016/S0925-5273(00)00063-3.

Pearl, Judea. 1988. “Probabilistic Reasoning in Intelligent Systems.” *Morgan Kauffmann San Mateo*. https://doi.org/10.2307/2026705.

Petropoulos, Fotios, Spyros Makridakis, Vassilios Assimakopoulos, and Konstantinos Nikolopoulos. 2014. “‘Horses for Courses’ in Demand Forecasting.” *European Journal of Operational Research*

237 (1): 152–63. https://doi.org/10.1016/j.ejor.2014.02.036.

Popović, Jovan B. 1987. "Decision Making on Stock Levels in Cases of Uncertain Demand Rate." *European Journal of Operational Research* 32: 276–90. https://doi.org/10.1016/S0377-2217(87)80150-9.

Rekik, Yacine, Christoph H. Glock, and Aris A. Syntetos. 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.: 182–94. https://doi.org/10.1016/j.ejor.2017.02.001.

Sherbrooke, Craig C. 2000. "Using Sorties vs . Flying Hours to Predict Aircraft Spares Demand." Virginia US.

Sterman, John D. 2000. *Business Dynamics - Systems Thinking and Modeling for a Complex World*. Boston: McGraw-Hill.

Weber, Philippe, and Lionel Jouffe. 2006. "Complex System Reliability Modelling with Dynamic Object Oriented Bayesian Networks (DOOBN)." *Reliability Engineering and System Safety* 91 (2): 149–62. https://doi.org/10.1016/j.ress.2005.03.006.

Weber, Philippe, Lionel Jouffe, and Paul Munteanu. 2004. "Dynamic Bayesian Networks Modelling the Dependability of Systems with Degradations and Exogenous Constraints." *11th IFAC Symposium on Information Control Problems in Manufacturing*. https://hal.archives-ouvertes.fr/hal-00131316/document.