Benchmarking spare part demand forecasting methods.

Daniel de Haan
Student 452942

Supervisor: Prof. dr. ir. R. Dekker

Second assessor: dr. V. Avagyan

Erasmus School of Economics
Erasmus University Rotterdam

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Master Thesis for programme: Economics and Business

Final version: December 2021

Abstract

The field of spare part demand forecasting focuses on methods that offer the optimal combination of reduced downtime and cost efficiency in a setting where demand may not be constant or stable. As newly developed methods enter the field, their ease of implementation and understanding needs to be put to the test. To check whether several methods would deserve more attention, this paper compares newer and well known spare parts demand forecasting methods on industrial and simulated data sets. The data sets are classified based on underlying data characteristics and the performance of each method is evaluated on forecasting accuracy and inventory control performance. The methods that are compared in this paper are Croston's method, Simple Exponential Smoothing (SES), the Syntetos-Boylan approximation (SBA), the Teunter-Syntetos-Babai method (TSB), Willemain's bootstrapping method, a machine learning method in the form of a Multi-Layer-Perceptron (MLP) and an approach based on the LightGBM Algorithm. SBA was the overall best performing method based on forecasting accuracy and the Willemain method was the overall best on inventory performance. The MLP and LightGBM methods were superior when extreme intermittency was present based on inventory performance, but their forecasting accuracy was lacking overall. These results were compared to existing literature and were found to be comparable. The main findings of this research indicate that the pre-processing of the data, the type of performance measure used and the type of data the method is applied to all appear to influence the performance of a method.

Table of contents

Chapter 1

Introduction and literature review

Keywords: Spare parts demand | Forecasting | Benchmarking | Industrial data vs. Simulated data | Intermittent demand | Inventory control | Programming in R

1.1 Introduction

Spare parts are essential in keeping machinery up and running when failure requires the replacement of vital components. Predictive and corrective maintenance complement these so called after-sales service replacements and further increase the importance of spare parts (Fortuin and Martin, 1999). This is why companies often keep stock of the most important spare parts to ensure that any downtime can be dealt with swiftly, because downtime is costly. However, keeping stock is also costly. Inventory holding costs can range from 5 - 45 percent of the cost price of the inventory per year, with an often used average of 25 percent (Durlinger and Paul, 2012). Not only are these holding costs of inventory relevant, component devaluation costs, price protection costs, product return costs and obsolescence costs may also be relevant (Callioni et al., 2005). This is why forecasting the demand is not only rather difficult, but also vital for correctly managing inventory (Willemain et al., 2004, Syntetos et al., 2015). The field of spare part demand forecasting focuses on methods that offer the optimal combination of reduced downtime and cost efficiency. Methods have also been devised to manage the uncertainty of demand in manufacturing (Bartezzaghi et al., 1999). As computing power has been steadily increasing over recent years and more attention has gone into spare part demand forecasting, newer methods have surfaced which may prove to be more suitable in certain situations. This leads to the main questions of this research:

Which type of method is best for which type of data?

and

Does the performance of the method depend on the performance measure used or on the data set to which it is applied?

In this paper several spare parts demand forecasting methods will be benchmarked on industrial and simulated data sets to layout their pros and cons. This will be done in five steps, where the first step will be familiarisation and implementation of all relevant published spare parts demand forecasting methods. After that, the industrial data sets used to evaluate the relevant methods will be gathered and made accessible through the use of Github. The next step will be to characterise these data sets and to subsequently generate simulated data sets with several common patterns that can be found in the industrial data sets. The forecasting methods gathered in step two will be evaluated on these data sets and then as the final step, the differences between the industrial and the simulated data sets will be identified and then related to the performances of the gathered forecasting methods.

1.2 Recent literature on spare part demand forecasting.

The literature of spare part demand forecasting can be divided into three main approaches, which each branch into several approaches: Time series methods, contextual methods and then studies which compare several methods, similar to this study. The framework for this substantive literature review is adapted from Pinçe et al. (2021). Time series methods are methods which use data from previous periods to forecast future data and they can be classified as either a parametric or non-parametric approach, based on whether the data is assumed to follow a known probability distribution (parametric) or if the distribution is assumed based on the data distribution (non-parametric).

Parametric methods are classified as either a modification of Croston's method, as an incorporation of demand obsolescence, as statistical bootstrapping, or as an approach that is not one of the previously mentioned ones but which is based on a parametric approach (Croston, 1972). On the side of the non-parametric approaches, the methods are classified as bootstrapping methods, the empirical method by Porras and Dekker (2008) and variations on the empirical method, and lastly neural network models. The third branch of time series methods focuses on improving the quality of forecasts through either classifying demand based on characteristics to categorize data or aggregating data to reduce demand variability and these two approaches are also referred to as forecast improvement strategies.

The collective of contextual methods can be divided into two distinct branches: Installed base forecasting and judgmental forecasting. Installed base forecasting investigates whether circumstances such as maintenance planning, the age of the equipment and other operating conditions influence the spare parts demand. Judgmental forecasting focuses on the opinion of experts on the effectiveness of forecasting strategies.

Subsequently, the third and for this paper most relevant category is the comparative studies category. These studies have provided comparisons between the most relevant methods for spare part demand forecasting at the time of their publication. Different studies have each used one or multiple industrial or simulated data sets to measure the performance of forecasting methods on different performance metrics. The most common performance metrics used are forecasting accuracy or inventory performance. This is why for this (comparative) paper, several spare parts demand forecasting methods will be benchmarked based on several industrial and simulated data sets. With the data sets, results and an overview of all the methods used available online, the results will be reproducible when newer methods are introduced into the field of spare part demand forecasting.

As an addition to the three main branches of the spare part demand forecasting literature, the M4 and M5 forecasting competitions and their main findings will be discussed (Makridakis et al., 2018, 2020a, Petropoulos and Makridakis, 2020, Makridakis et al., 2020b). This international forecasting competition puts competitors to the test on a uniform task meant to test their forecasting method on equal parameters and performance metrics. The aim of the competition is to advance the field of forecasting and look for new ways to apply known or newly developed techniques.

1.2.1 An overview of time series methods

Parametric approaches

The first set of time series forecasting methods are the parametric approaches, where one assumes that demand can be explained because it follows a certain distribution. The aim of these approaches is to gather information on the historical mean and variance of demand and use this to predict future demand. However, in a situation where demand is very low or zero for extended time periods, predictions often result in inaccurate estimates. Especially classical techniques such as Simple Exponential Smoothing (SES) fall short in predicting demand in these situations (Pinçe et al., 2021). This type of inconsistent demand can also be called intermittent or erratic demand. Croston (1972) was the first to notice this and in his 1972 paper he developed a new method, now referred to as *Croston's method* or *Croston*. Croston figured out how to solve the problems that SES has with intermittent demand. His method split the demand estimate into two parts: the inter-demand time and the demand size. These separate parts are forecast individually using SES, resulting in more accurate and smoother estimates. Croston has since it's development been used as a performance benchmark, because of it's smooth estimates with less error variation than SES. The result is that Croston provides the same service level at lower safety stocks. Lower safety stocks mean a lower average stock value, resulting in lower average inventory holding costs.

Next to intermittent demand, obsolescence is also an issue. Obsolescence refers to when the demand of an item reduces (potentially to zero) over time. When demand is intermittent, obsolescence may not be detected because zero demand periods are common and thus do not stand out.

Syntetos and Boylan (2005) developed a method which is similar to Croston because it also splits the demand and this method is called the Syntetos-Boylan Approximation (SBA). In their variation of Croston's method, a bias correction coefficient is introduced. This bias correction means that SBA provides more accurate results for intermittent demand. This is why SBA is often used alongside Croston as a benchmark (Syntetos and Boylan, 2006, Teunter et al., 2011, Zhu et al., 2017, Babai et al., 2019). In general, Croston outperforms SBA in terms of service level and SBA outperforms Croston in terms of forecasting accuracy and the differences depend on the demand patterns (Pinçe et al., 2021). These differences will be empirically tested on simulated and industrial data sets later in this paper. As mentioned earlier, when obsolescence or gradually decreasing demand is present, SES, Croston and SBA may under-perform. For these specific demand patterns, TSB was developed by Teunter et al. (2011). This new method also uses the demand size forecast, but TSB combines this forecast with the demand probability forecast instead of the inter-demand interval forecast. This method also updates the demand size when demand does not occurs. This means that a forecast is adjusted downward when there is no demand, decreasing the time needed to spot obsolescence. Initially TSB was shown to have better accuracy than SES, Croston and SBA, but when empirically tested by Babai et al. (2014), the performance was not considerably better than the other methods. Because of these results, Babai et al. (2019) propose a new method, called modified SBA. This method can be seen as a mixture of the positives of SBA and TSB, where the forecast updates are more towards TSB when the risk of obsolescence increases. Then lastly, Hyperbolic Exponential Smoothing (HES) is introduced by Prestwich et al. (2014). HES also decays forecasts when demand is not present, but hyperbolically instead of exponentially like with TSB. They showed that TSB is more accurate in estimating, but that HES is more straightforward in practical applications.

Lastly, some dynamic time-series methods have been developed. First, Pennings et al. (2017) introduce a method which aims to anticipate incoming spare parts demand by incorporating positive cross-correlation between inter-arrival times and demand sizes. This proposed method outperforms SES, Croston, SBA, TSB and Willemain et al's method when the demand size and the inter-arrival time are strongly positively cross-correlated. Subsequently, Snyder et al. (2012) introduce a method which incorporates possible random shifts in the demand distribution mean by smoothing the previous period's mean and demand realizations through the use of Poisson, negative binomial and hurdle shifted Poisson distributions. The results show considerable accuracy gains, however, their complexity means that they require more advanced knowledge to perform and provide less intuitive conclusions. Then finally, Jiang et al. (2020) propose a method aimed at reflecting the changes in the underlying demand process with a mixed zero-truncated Poisson hurdle model. The proposed method showed interesting results on electric power company data, however, as with the method by Snyder et al. (2012), this method needs additional performance assessment on both industrial and simulated data sets.

Parametric, nonparametric and smooth bootstrapping

When a data set lacks data or shows patterns which are hard to incorporate in a parametric approach, bootstrapping may be an option. Bootstrapping can be described as simulating data by looking at the available data and then filling in the gaps, providing the researcher with a larger body of data with more predictable parameters. This was first introduced by Efron (1979). The generation of data is repeated multiple times to simulate a usable distribution within the data. Because the necessary data or data structure to test whether a method works may not always be available in real life, this additional data creation or adaptation allows the researcher to test their method on situations which may prove difficult when encountered in real life situations. Unfortunately, the demand is not always distributed parametrically. If the lead-time demand can not be described clearly by a parametric distribution, parametric methods may not perform well. This is why other methods have been developed that take a nonparametric approach. These methods try to figure out empirically how the lead-time demand can be described and thus how it can be forecast. According to Smith and Babai (2011) this means that nonparametric methods can be applied to a greater variety of data than parametric methods, such as data with highly erratic demand. Bookbinder and Lordahl (1989) and Hasni et al. (2019) agree that using a nonparametric approach also removes the possibility to assume the incorrect distribution, which could result in incorrect estimates.

The current body of literature on bootstrapping methods for forecasting spare parts demand has different approaches. There are parametric, nonparametric and smooth bootstrapping methods. Smooth methods are different in that they add a random amount from typically the uniform or normal distribution to each sampled data point, in order to provide a discrete empirical function when the actual distribution appears to be continuous (Hasni et al., 2019). The first set of parametric methods are Snyder's parametric bootstrapping methods. Snyder (2002) adapted Croston's method and SES to integrate demand forecasting with inventory control and these methods are referred to as the log-space adaptation (LOG) and the adaptive variance version (AVAR). For a more in-depth explanations of these and all of the following bootstrapping methods, consult the overview in Hasni et al. (2019). One of the initial parametric bootstrapping methods is from Bookbinder and Lordahl (1989). Their method performs best when the lead-time demand distribution is assumed to be neither normal or lognormal, because then other parametric bootstrapping methods are superior. Their method can be seen as an extension of Efron's classical introduction of bootstrapping with two extra steps: The mean and the standard deviation of the lead-time demand values from the bootstrap are calculated and a theoretical density function is obtained based on the empirical mean and standard deviation of the lead-time demand values.

This adaptation means that this method performs well on service levels and lowers costs when the lead-time demand distribution assumption holds (Hasni et al., 2019).

The next and probably most used bootstrapping approach is the one by Willemain et al. (2004), also referred to as WSS. Their method can be considered as both a smooth bootstrap because they used a jittering process on their demand predictions and as a parametric bootstrap because this jittering process assumes a normal distribution (Hasni et al., 2019). The main problems that WSS aims to solve are: autocorrelation, frequent repeated values and relatively short series. The method uses a Markov model to generate zero and non-zero values over the forecast horizon based on historical demand, after which every non-zero state marker is replaced at random by a randomly sampled demand size from historical demand sizes. After this, the jittering process is initiated, which creates demand size values which are not observed historically and thus intend to decrease lumpiness in the data. A generalisation of WSS was made by Kocer (2013) by introducing higher order Markov chains.

The next nonparametric bootstrapping methods are those by Wang and Rao (1992) and Viswanathan and Zhou (2008). Wang and Rao apply Efron's bootstrap twice to create an inventory system focused on the reorder point to estimate the lead-time demand distribution and estimate the reorder point. The approach of Viswanathan and Zhou, also referred to as VZ, constructs the lead-time demand distribution by using bootstrapping to create the demand intervals and use sampling to generate demand sizes. A slightly different method is the empirical method of Porras and Dekker (2008). This nonparametric method is simpler because is does not require sampling and the lead-time demand data is generated from the data by evaluating historical demand based on blocks equal to the length of the lead time. Several followup studies have been done to assess whether changes to this method provide better results. Boylan and Babai (2016) found that instead of assessing the lead time blocks separately, they should be analysed with overlap, especially when lead times are longer. Van Wingerden et al. (2014) extended Porras and Dekker's method by randomizing the lead time and found that their proposed adaptation performed better on certain performance metrics. Similarly, Zhu et al. (2017) found that including extreme value theory to assist in predicting possible extreme values improves the method and could lead to higher service levels, but that this adaptation performs poorly when there are few demand points over many periods (a high degree of intermittency).

Neural network methods

The following set of methods apply a form of machine learning in their approach, usually in the form of a neural network algorithm application. The machine learning methods used for forecasting spare parts demand can be classified as supervised learning methods, as the variable of interest is known and the methods simply aim to unearth the underlying dependencies within the demand data and use these to predict future values (Molnar, 2020). In other applications, neural networks are often criticised because the interpretibility of the underlying dependencies becomes less clear than with other methods, but since spare parts demand forecasting focuses on results (forecasting accuracy or inventory performance), neural networks seem to suffer less from this drawback. Gutierrez et al. (2008) were one of the first to apply a neural network approach to the field of spare parts demand forecasting, comparing their method to SES, Croston and SBA. The results showed that forecasting accuracy was relatively good, but that sufficient data has to be present to train the neural network. An adapted version of their method was introduced by Mukhopadhyay et al. (2012) which outperforms the traditional methods in their comparison and a generalisation by Kourentzes (2013) showed that neural network methods are worse at predicting than traditional methods, but offer a superior service level. Lolli et al. (2017) introduce another neural network approach and compare it to all three of these methods and find that their method works faster and easier, but that the methods by Gutierrez et al. (2008) and Mukhopadhyay et al. (2012) perform better in general. Lastly, Guo et al. (2017) created an ensemble, combining a neural network, exponential smoothing and hierarchical forecasting to create a combination that is superior at forecasting compared to applying the methods individually.

Forecast improvement strategies

Forecast improvement strategies comprise the last set of time series methods, with the methods being either a form of demand classification or data aggregation. With demand classification, the researcher attempts to classify the demand based on distinct characteristics in order to match those characteristics to the approach that they intend to use. Williams (1984) was one of the initial articles on demand classification and demand is categorized based on the pattern and then the categories are matched to a distribution that fits the specific pattern for each demand period. Johnston and Boylan (1996) provide support for a theory by Willemain et al. (1994) that SES should be superior to Croston for data with very high or low intermittency and introduce a method that uses the variability of the demand size and the average inter-demand interval to forecast expected demand which supports this theory. In their article they identify a demand classification parameter in the average inter-demand interval (p). Syntetos and Boylan (2005) suggest that the squared coefficient of variation of demand (CV^2) should also be considered as a demand classification parameter and they find cutoff values through an analytical comparison to be used in their demand classification scheme. These cutoff values are criticised by Kostenko and Hyndman (2006) and they propose different cutoff values for (p) and (CV^2).

To elaborate on the earlier works, Boylan et al. (2008) investigate the stock-keeping capabilities of the demand classification scheme and find that the amount of periods with zero demand can be a demand classification parameter. This line of thought is further elaborated upon by Syntetos et al. (2011) by recommending heuristic rules which can be used to pick a theoretical demand distribution that is best for inventory control. An extensive empirical analysis on different issues regarding demand classification is introduced by Lengu et al. (2014)

and they develop a demand classification scheme based on order size. Using the Kolmogorov Smirnov (K–S) goodness-of-fit test, Turrini and Meissner (2019) attempt to find the best fitting distribution for their data and provide thoughts on how the goodness-of-fit relates to inventory performance for demand classification schemes. Focusing on the determinants of forecasting accuracy, Petropoulos et al. (2014) shed light on how researchers should determine their forecasting method. They find that forecasting accuracy is mainly influenced by the cycle and the randomness of demand and that accuracy decreases when the forecasting horizon is increased for fast-moving data and by the inter-demand interval for intermittent data. They also find that increasing the length of a series slightly increases forecasting accuracy for all types of data.

Data aggregation is another form of improving forecasting accuracy, this time through the aggregation of data based on similarity in demand pattern. For example, Willemain et al. (1994) showed that examining data on a weekly basis instead of a daily basis already increases accuracy when using Croston to forecast. Nikolopoulos et al. (2011) attempt to further narrow down an optimal aggregation level and empirically investigate how aggregating demand equal to the lead-time length could increase forecasting accuracy. Their results show that forecasting accuracy may be increased and forecasting variance may be decreased through data aggregation, specifically using their Aggregate-Disaggregate Intermittent Demand Approach (ADIDA). Babai et al. (2012) elaborate on the ADIDA method by validating it's inventory performance in combination with Croston, SBA and SES and the results show that ADIDA improves service levels across all methods. Mohammadipour and Boylan (2012) show that aggregation provides more accurate forecasts when applying their temporal aggregation scheme, they do however indicate that the effects of aggregation on inventory performance has yet to be investigated. Petropoulos et al. (2016b) provide an alternative look on ADIDA and introduce a new framework called iADIDA (inverse ADIDA). They empirically test forecasting accuracy and stock-control and find that their approach works specifically well for the sections of time series data with the highest data-volume variance and they argue that for the other sections of data the forecasting method should be chosen according to the findings of Petropoulos et al. (2014), which would be SBA or TSB in most cases. Boylan and Babai (2016) assess whether the temporal aggregation needs to be done with or without overlapping the time buckets and they found that overlapping generally outperforms the non-overlapping approach. These results are in line with their opinion on how the overlapping of time buckets when using the empirical method by Porras and Dekker (2008) increases forecasting accuracy.

Temporal aggregation is not the only form of data aggregation. Data can also be aggregated based on other characteristics to form groups of items. Moon et al. (2012) show that grouping items may reduce forecasting error and decrease inventory costs. Li and Lim (2018) introduce a method called greedy aggregation–decomposition (GAD) which uses both top and bottom aggregation levels in a hierarchical forecasting framework. The method shows superiority to the methods it is tested against, such as SBA, Croston, TSB and other data aggregation methods. Other forecast improvement strategies include combining forecasts derived from alternative methods (Petropoulos and Kourentzes, 2015) or improving outlier detection (Zhu et al., 2017, Romeijnders et al., 2012). However, Pinçe et al. (2021) argue that demand spikes should simply be treated as outliers and should not be included in the forecasting, as they indicate predictive maintenance, which can be anticipated and acted accordingly upon.

Conclusion time series methods

Overall it can be concluded that time series methods comprise of many different approaches and methods, all using their own approach to capture the variability of spare parts demand forecasting. Often Croston, SBA and other traditional methods are used as a benchmark because of their generally solid performance. Balancing between forecasting accuracy and inventory performance, obsolescence of items, finding the right distribution (if any) of the lead-time demand and also being able to perform on both industrial and simulated data sets are all problems these methods face and attempt to overcome in their own way. It is then also unsurprising that results appear to be dependent on the data and the method used and that performance might be severely influenced by the performance measure used. The general conclusion from the literature thus far seems to be that each method has it's own perks, performing specifically well in very specific situations. This line of thought leads to the idea that a combination of methods that could capture the superiority of each method while still attempting to keep implementation costs low could result in a generally superior approach.

1.2.2 An overview of contextual methods

All of the previously discussed approaches used historical data to attempt to predict the lead-time demand, but this means that adapting to significant underlying changes might be slow. Other shortcomings of time series methods are that historical data needs to be sufficiently present to be able to make accurate predictions about the lead-time demand and that changes in maintenance schemes are passively understood instead of proactively accounted for (Wang and Syntetos, 2011). This is especially unfortunate when the researcher actually has knowledge on the nature of the data and how it is likely to change in the future, meaning that a forecasting approach could likely be improved by the researcher's contextual knowledge. If the researcher already has this knowledge ahead of demand, this can be referred to as advance demand information (ADI) (Zhu et al., 2020). This is why in practice, statistical methods are often guided by the knowledge and judgement of the researcher (Sanders and Ritzman, 1992). A practical example can be seen in the fashion industry, where the knowledge of customer taste and upcoming trends is vital for accurately predicting new product success (Seifert et al., 2015). The application to spare parts demand forecasting is however not as subjective, but rather about understanding how far along the items or groups of items are in the life-cycle of the equipment the spare parts are installed in, which may be clear from information outside of historical data. An example could be that the manufacturer of the original equipment is promoting a newer version of their product, which may result in spare part obsolescence for the older version. To ensure that the contextual information is somehow incorporated in the forecasting method, contextual approaches were developed and these methods have gained more attention from researchers in past years. Contextual methods are divided into judgemental forecasting and installed base forecasting approaches.

Judgemental forecasting

It is widely understood and observed in practice that researchers make judgemental adjustments to statistical forecasts by making use of managerial knowledge (Klassen and Flores, 2001, Goodwin, 2002, McCarthy et al., 2006, Fildes and Goodwin, 2007, Makridakis et al., 2008, Boylan and Syntetos, 2010) and while some empirical studies have shown that the effects of these adjustments may improve statistical forecasts (Turner, 1990, Mathews and Diamantopoulos, 1986, 1992), other studies suggest that this may vary (O'Connor et al., 1993, Sanders and Ritzman, 2001, Sanders and Manrodt, 2003, Franses and Legerstee, 2010, Petropoulos et al., 2016a). Research has however been limited on the integration of judgement in spare parts demand forecasting, and even more so for intermittent demand specifically (Pinçe et al., 2021). Syntetos et al. (2009) investigate this integration through an empirical analysis of a pharmaceutical company's demand forecasts. The pharmaceutical company uses common statistical software to predict demand and then uses managerial insight to adjust the predicted demand. Syntetos et al. (2009) find that negative adjustments are more effective in resulting in accurate predictions than positive adjustments and that larger negative adjustments perform particularly well. They also suggest that the relatively poor performance of positive adjustments may be because the forecaster was biased by managerial pressure, an optimism bias or by the intention to be favoured by suppliers. Another look on the integration of expert opinion in forecasting for spare parts demand is provided by Boutselis and McNaught (2019). They specifically analyse military operations spare parts forecasting in the final period of a military operation. In these situations, predicting how much demand is required for the final period accurately may result in significant logistical advantages. Their findings suggest that expert adjustment is often made in the wrong direction and that a SES forecast without expert judgement was getting better results. Their proposed Bayesian Network (BN), particularly the version integrated with machine learning, was showed to outperform forecasts with expert adjustment and a logistic regression. Besides these papers, judgemental adjusting is more commonly researched in a supply chain management setting, which is why it will not be investigated further for the purpose of this paper.

Installed base forecasting

As previously mentioned, making adjustments to statistical forecasts is common in practice. The type of information that the researcher uses to adjust the forecast may however also be intrinsic to the spare part and it's installation environment instead of judgemental or managerial knowledge. This type of information can be referred to as *installed base information* (Borchers and Karandikar, 2006, Dekker et al., 2013) and the *installed base* has several interpretations in the literature, such as the total amount of customers using a specific part or the total amount of units in use of a specific part (Borchers and Karandikar, 2006). Commonly tracked characteristics include spare part location, current owner and user, the application, the operating environment, product status and service history (Ala-Risku et al., 2009). Installed base methods make use of the installed base information to aid in predicting spare part failure and increase forecasting accuracy (Pinçe et al., 2021). The added value of this information has been investigated by Jalil et al. (2011) in the planning of spare parts demand and they find that there are potential gains in using installed base information. In order to maximise those gains, they suggest that the researcher should aim to understand how data errors may occur in practice and align the business environment with installed base data usage.

Dekker et al. (2013) provide a review on the utilisation of installed base information in industry practice. They found that installed base information usage in forecasting was already being mentioned in the forecasting of new product adoption by Brockhoff and Rao (1993), that Cohen et al. (1990) theoretically mentioned the application to spare parts and that Auramo and Ala-Risku (2005) discussed installed base information for service logistics. Dekker et al. (2013) conclude that installed base data can be used to improve forecasting accuracy, even when the quality of the data may not be optimal, compared to only using historic demand for forecasting. But these and other early papers such as Petrović and Petrović (1992), Ghobbar and Friend (2002) and Aronis et al. (2004) do not benchmark their proposed installed base methods against the traditional or other spare parts demand forecasting methods, with Hua et al. (2007) being one of the first to do so (Pinçe et al., 2021). Their developed approach takes plant and equipment overhaul situations as explanatory variables for the prediction of the lead-time demand through regression and they compare their method to SES, Croston and an adaptation of Willemain's bootstrap, resulting in their method outperforming the traditional methods. Similarly, Wang and Syntetos (2011) compare a maintenance driven approach integrating failure timing to the SBA approach. Their results show more accurate forecasts in almost all cases. Similar results are produced by Romeijnders et al. (2012), who show that the forecasting error can be reduced significantly if planned maintenance schemes are taken into account when forecasting. In a more recent study by Zhu et al. (2020) an approach combining forecasting aided by maintenance planning and an inventory control method based on the forecasts. They find that major inventory cost savings can be attained when the method is compared to the traditional time series forecasting methods. Another major advantage of their method is the limited data requirement compared to time series methods.

Conclusion contextual methods

Contextual methods use knowledge outside of historical data to improve forecasting accuracy. Managerial knowledge and judgement are used in judgemental forecasting and installed base forecasting makes use of known product or maintenance characteristics known as installed base information to more accurately predict spare part demand. Overall contextual methods appear to positively influence forecasting accuracy, as long as the influence of the contextual knowledge is based on data or set information, as acting based on intuition generally influences the forecasting negatively. Other advantages of using contextual methods over time series methods are the lesser need for extensive historical data and the more proactive nature of the approaches.

1.2.3 An overview of performance comparisons

There have been several papers like this one that intend to compare forecasting methods or a specific subset of these methods used in practice. In order to objectively compare the proposed methods, performance benchmarks are used. For the field of spare parts demand forecasting, the most important performance benchmarks are either based on the predictive accuracy of the method or the method's inventory control performance (Pinçe et al., 2021). When the predictive accuracy is the main focus, the performance benchmark indicates how close the predictions made by a proposed approach are to the known demand, often tested by splitting the data into a test and training data set. If the difference is large, the method will score poorly on the forecast accuracy performance benchmark. For the inventory performance benchmarks, objectives such as inventory holding costs, service levels, reducing stockouts and reducing the overall on-hand inventory relate to a method's performance. After reviewing the most common performance benchmarks, some recent comparative papers will be discussed.

Forecasting accuracy

As the accuracy of a prediction determines it's usefulness, determining the accuracy objectively requires a forecasting accuracy measure that is universally applicable and that does not benefit one of the benchmarked methods more than another. There are two main types of forecasting accuracy measures, *relative* and *absolute* forecasting accuracy measures. Relative accuracy measures compare a method to another method or to a baseline accuracy, while absolute accuracy measures give an indication of the forecasting error (Syntetos and Boylan, 2005). This forecasting error can be interpreted as an indication of the deviation of the predicted demand from the actual demand. Next to the relative and absolute accuracy measures, there are also studies using a simpler comparison in which they measure the amount of times a method performs better or best when compared to the other methods in an empirical setting. Lastly, there are measures developed specifically for a method or application, such as the forecasting accuracy measures introduced by Willemain et al. (2004), Hua et al. (2007) and Kim et al. (2017). In their broad comparison, Pinçe et al. (2021) find that the most commonly used forecasting accuracy measures are absolute accuracy measures and the authors attribute this to their simplicity in use and interpretation.

Inventory control

Although accurate forecasts may lead to correctly predicting the demand, they may not always lead to a desirable inventory performance. Because of this, Syntetos and Boylan (2006) state that a forecasting method should be judged based on the inventory performance in order to measure it's effectiveness. Similarly, Teunter and Duncan (2009) find that forecasting accuracy measures are ineffective when applied to intermittent data and instead suggest comparing the methods based on inventory management implications and service level. Syntetos et al. (2010) further elaborate on Syntetos and Boylan (2006) with an empirical experiment and find that very minor changes in forecasting accuracy may have major implications on stock management. Including both in a comparison could thus result in major cost implications compared to only using forecast accuracy as a performance metric. Pinçe et al. (2021) find that the most used inventory control performance measures in recent spare parts demand forecasting literature are the service level and tradeoff curves. They also find that most studies use a combination of inventory performance measures to more elaborately assess the performance of an approach.

An insight into comparative papers

As the spare parts forecasting methods have developed, papers comparing their performance have also surfaced more frequently. One of the first comparative papers by Willemain et al. (1994) compares Croston and SES. In their comparison they used data that was simulated in such a way that it violates the normality and independence assumptions from Croston (1972). Even in these less than ideal situations, Croston outperformed SES both on this simulated data and later on industrial data. Another study by Sani and Kingsman (1997) compared the performance of five approaches. They found that a simple moving average is best, followed by Croston, and that both are superior to SES in terms of inventory performance for intermittent or low demand. Syntetos and Boylan (2006) also find that moving average performs well overall, but in their comparison paper SBA outperforms moving average in terms of inventory performance. Another broad comparison of thirteen methods is performed by Ghobbar and Friend (2003), using the Mean Absolute Percentage Error (MAPE) as the performance indicator. In their comparison, weighted moving average is the superior method. However, Teunter and Duncan (2009) critically observe that choosing the performance measure greatly influences the success of each approach. Through an empirical example they show that forecasting zero demand in each period leads to a better performance compared to the traditional methods when the standard forecast accuracy measures are used. Forecasting zero demand would of course result in very poor inventory performance, showing the ineffectiveness of using the wrong performance metric. Using inventory performance, they also find that SES and moving average are consistently outperformed by Croston, SBA and a parametric bootstrapping method. Pinçe et al. (2021) find that most comparative papers do not include bootstrapping methods and that Syntetos et al. (2015) and do Rego and De Mesquita (2015) are some of the few that do. Syntetos et al. (2015) find that the bootstrapping method by Willemain et al. (2004) does outperform the traditional methods in most situations, but that the traditional methods are simpler in use and interpretation. do Rego and De Mesquita (2015) find that the bootstrapping approach by Zhou and Viswanathan (2011) is best for lumpy demand and that SBA is best for erratic demand.

To synthesize the spare parts demand forecasting research, Pinçe et al. (2021) perform a quantitative literature analysis on 53 papers reporting method comparisons. They compare the methods used in each paper by counting the amount of times each method outperforms the other methods, resulting in better performance scores. These better performance scores are then later divided by the total number of comparisons to give a percentage better score, which indicates how often the method was superior. This score is then averaged to give an overall performance indicator for each method, which is called the Average Percentage Better (APB) score. They find that Croston is outperformed by SBA in 85,7% of the papers in which they are compared when it comes to forecast accuracy measures. On inventory measures, the results are less clear, but the conclusion is that Croston generally outperforms SBA for intermittent or erratic demand and that Croston offers slightly better service levels. They also provide an extensive comparison of performance results for traditional methods, newer parametric and nonparametric approaches, bootstrapping methods and approaches including contextual information or data aggregation. Their main conclusions are that inventory performance should be focused on more than forecasting accuracy due to practical relevance, that judgemental adjustments seem to always positively influence forecasts and that data-intensive methods such as neural networks show great potential, if implementation costs can be minimised. A limitation of this research is that they did not apply any of the methods themselves.

Performance comparison conclusions

Comparing methods' performance is nothing new. Usually a comparison is made with one of the standard spare parts demand forecasting methods, such as SES, Croston and SBA because of their applicability to intermittent demand. Recent studies also showcase the importance of choosing the correct performance metric between the available forecasting accuracy and inventory performance measures. Overall the consensus seems to be that there is not one particular performance measure that is suitable for every method and every type of data. This is especially apparent in the difference in performance on industrial and simulated data sets. A relevant and feasible performance measure should be chosen based on the variability of the demand sizes and the intermittency of demand.

1.2.4 The M competitions

To advance the theory of forecasting and provide an equal testing ground for forecasting methods, the M competitions were initiated by S. Makridakis and colleagues, with the first (M1-competition) being held in 1982 (Makridakis et al., 1982). As the most recent two M competitions showed some promising approaches for the field of spare parts demand forecasting, the M4 and M5 competitions and their methods will be reviewed in this section. As the implications and results of the first three M competitions were used and incorporated in the approaches in the M4 and M5 competitions, these will not be reviewed separately. For details on the first three M competitions, see Makridakis et al. (1982), Makridakis et al. (1993) and Makridakis and Hibon (2000).

The M4 competition (2018)

In the M4 competition participants were challenged to predict 100.000 series and 100.000 Prediction Intervals (PIs) (Makridakis et al., 2018). As with all of the M competitions, the aim of the M4 was to further the field of forecasting, but this version also focused on three things: Increasing the number of series compared to the previous editions, not only assessing the point forecasts but also the PIs and to include machine learning approaches. As the predictions made by participants have to be benchmarked against a method, ten standard methods were introduced, of which Comb had been chosen. Comb is described as a combination approach in which the arithmetic average of the Simple, Holt and Damped exponential smoothing models is calculated. Comb showed to be simple and easy to implement and it showed the highest accuracy based on the Symmetric Mean Absolute Percentage Error (sMAPE), one of the forecasting accuracy performance measures used. The main performance measure used for deciding the winner is the Overall Weighted Average (OWA) of the sMAPE and the Mean Absolute Scaled Error (MASE). On the OWA, 17 methods showed an improved forecasting result compared to Comb. The M4 did not consider inventory performance.

One initial interesting result is that the approaches applying only a machine learning method introduced in the M4 did not perform particularly well, with all of them performing worse than the Comb benchmark. Makridakis et al. (2018) contribute this to the fact that the application of machine learning to forecasting was still in its infancy and that the methods also require a lot of computational power. This could have meant that machine learning methods were not able to finish their submission in time for the submission deadline set by the M4 competition. The method that performed the best overall did however incorporate a recurrent neural network (RNN) model in the approach, which is a form of machine learning. The method can be described as a hybrid approach because it mixes the aforementioned RNN with exponential smoothing formulas. Another innovative aspect of the winning method was that it not only used the whole data set, but also used the individual series for the predictions, making the method significantly more accurate. Overall, the main takeaway for forecasting from the M4 competition is clear. Approaches utilizing the best of different methods to combine them into one better method show promising results, especially when the weights of the different methods are determined through machine learning.

The M5 competition (2020)

Following the M4 competition, several adjustments were suggested by various commenters and the design of the M5 competition was aimed at addressing these issues (Makridakis et al., 2020b). Some important adjustments were made compared to the M4. The initial change was that the competition was hosted by Kaggle on this occasion. Kaggle is an online platform where data scientists, practitioners of machine learning and other data enthusiasts share different methods and challenges to learn more about all the various aspects of data science, including forecasting (Kaggle, 2020). This meant that competing was easier, resulting in more submissions. Another big change was the addition of contextual variables, which could be used to improve the accuracy of the forecasts. These contextual variables were special events and holidays, selling prices and the presence of special promotions aimed at lower income families as a binary variable. The third change was the structure of the data. This time the data more closely resembled real-life data because the time series were grouped and correlated and they were organized in a cross-sectional structure. This new data also displayed intermittency, which makes the results of the M5 competition more relevant to spare parts demand forecasting than the previous M competitions. In total, the data consisted of unit sales at Walmart over the course of approximately five and a half years, across various locations and product categories. This variation of dimensions provides the competitors with various ways to group the unit sales.

The performance measure used for the M5 competition was also slightly different compared to the previous iteration, but the MASE was again used to produce the final forecasting accuracy performance measure. The variant used in the M5 competition is the Root Mean Squared Scaled Error (RMSSE), which is more suitable for the intermittent data. After calculating the RMSSE, the average of all of the RMSSE results across all series are weighted and this Weighted RMSSE (WRMSSE) is the final overall accuracy measure. In the M5 competition, special attention was also given to the benchmarks against which the methods would be measured. The benchmarks were selected based on several factors such as popularity and computational requirements (Makridakis et al., 2020b).

The results of the M5 competition showed much larger improvements compared to the best benchmark method than the earlier competitions, with five methods showing an improved forecasting accuracy of more than 20%. The improvement of the winning method was 22.4%. These improvements were however not equal amongst different aggregation levels, which ranged from the total amount of observations as one level all the way down to every observation. For the M5 data set this meant that aggregation levels 2-5 were aggregated unit sales per category, State, store or department for all products combined, and lower aggregation levels continue to increase the size of the amount of series. For example, the winning method performed best at four aggregation levels, namely 3, 7, 8 and 9, but did not win any of the other eight. The other aggregation levels were won by other methods and the winning method won on having the lowest overall WRMSSE. This suggests that in practice, finding the best forecasting method may also require investigating which aggregation level to choose, as this influences the accuracy. Amongst the top five methods, four used a variation of LightGBM, which is a machine learning algorithm. This algorithm can be used to develop an approach fairly easily, as the algorithm is very forgiving in the required input, is low on computing power and requires little optimization of features and data, although the customization is near limitless. LightGBM was also used in multiple other forecasting competitions and the winners of those competitions also based their approaches on LightGBM (Makridakis et al., 2020b).

Overall, the M5 competition provides the field of forecasting with several findings. Firstly, machine learning methods have evolved to be potentially superior to simpler methods. In the M4 competition, the machine learning methods were flawed in that they required a lot of data cleaning, pre-processing and computational power. In the M5 competition however, all of the top five methods used a form of machine learning, suggesting machine learning may now be better than statistical benchmarks and the simpler methods that could compete in previous competitions. The second main finding is that an approach based on a combination of methods is, similarly to the M4 competition winner, seems to be the superior option. This takes advantage of all the positives from each method to create an approach that is better than when the methods are applied individually. The third finding was caused by the way the data was provided to the participants in this version of the competition. Because the series were correlated this time and could be chronologically aligned, cross-learning was not only part of all the top methods, it made computing power requirements lower and allowed the methods to learn from all the information that the data set had to offer. Furthermore, the M5 competition showed that effective cross-validation, adjusting the forecasts based on prior knowledge and using the available contextual variables to aid in the forecasting all helped in creating more effective forecasting methods. All of these findings suggest that machine learning approaches tailored to the aggregation level and data set may be the best way forward for forecasting in practice (Makridakis et al., 2020b).

M competitions conclusions

The M competitions provide an equal testing ground for the most advanced methods in forecasting. By analysing what each method does well, the field of forecasting and forecasting in practice can be improved upon. The M4 competition showed that machine learning was on the rise, but that there were still flaws that made a combination of statistical methods better. In the M5 competition the machine learning methods improved so much compared to the M4 that almost all of the top methods applied a form of machine learning in their combined approach. Especially the LightGBM algorithm showed promising results for many of the competitors. As the M competitions focus on forecasting in general and not on spare parts demand forecasting, the takeaways for this paper come mainly from the M5 competition, as the data set used in that competition showed intermittency. Because the machine learning methods showed such superiority in this competition, to evaluate the applicability to spare parts demand forecasting, a machine learning method applying the LightGBM algorithm will be compared with some of the parametric and nonparametric methods highlighted earlier in the literature review and the methodology applied for the comparison will be elaborated on in the next section of this paper.

Chapter 2

Methodology

To clearly show the process of this paper, in this section first the experimental design will be graphically shown, then the selected methods will be introduced and technically described, followed by an elaboration on the used forecasting accuracy measures and inventory control measures. After that the demand classification methodology based on the four categories from Boylan et al. (2008) will be elaborated on.

2.1 Experimental design

The experimental design is based on the research questions introduced in the introduction of this paper, which revolve around the different results one may encounter when applying spare parts demand forecasting methods to different (types of) data sets. The experimental design is graphically represented in 2.1. First both the industrial and simulated data sets have been gathered. Some data wrangling is required for the industrial data sets and then both the industrial and simulated data sets are further explored and the demand is classified in the data section of this paper. After the classification, the selected methods will be applied, after which the methods are compared based on forecasting accuracy, inventory control measures and the differences between the results caused by the difference in the data sets is discussed.

2.2 Selected methods for the comparison

As one the initial methods introduced in the field of spare parts demand forecasting and because it is commonly used as a benchmark, the first method to be considered in the comparison is Croston's method. Predictions made by Croston are based on two separate components, the demand size z_t (which has to be non-zero) and the inter-demand interval p_t. z_t has to be non-zero in at least two periods because the predictions are updated only when demand occurs. The

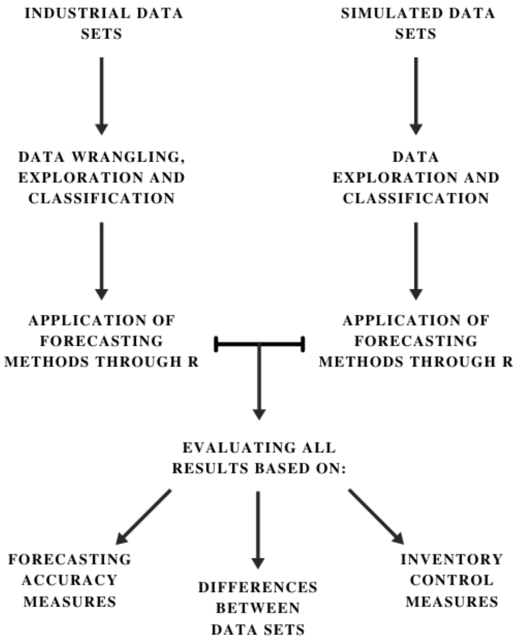

INDUSTRIAL DATA SETS
SIMULATED DATA SETS
DATA WRANGLING, EXPLORATION AND CLASSIFICATION
DATA EXPLORATION AND CLASSIFICATION
APPLICATION OF FORECASTING METHODS THROUGH R
APPLICATION OF FORECASTING METHODS THROUGH R
EVALUATING ALL RESULTS BASED ON:
FORECASTING ACCURACY MEASURES
DIFFERENCES BETWEEN DATA SETS
INVENTORY CONTROL MEASURES

Fig. 2.1 The flow of the experimental design.

predictions from Croston are given by the formula

$$\hat{y}_t = \frac{\hat{z}_t}{\hat{p}_t}. \tag{2.1}$$

The initial observation of the series is used as the initial value for the predictions and both z_t and p_t are predicted using SES with a smoothing parameter optimized by a cost function, as advised by Kourentzes (2014). The final output from Croston is the average estimated demand for each time period in the forecasting horizon.

For Croston and the following three methods (SES, SBA and TSB), the "tsintermittent" R-package by Kourentzes (2014) was used. This package requires the selection of several parameters, one of which is a cost function, to optimize the methods. All of the parameters are discussed extensively by Kourentzes (2014) and the options for the cost function are the mean squared error (MSE), the mean absolute error (MAE), the mean squared rate (MSR) and the mean absolute rate (MAR). The conclusion from their empirical research is that the MAR cost function is superior for every method apart from TSB. The cost function they find most suitable for TSB is the MSR cost function. The cost functions used in this paper are therefore according to their findings: the MAR cost function for Croston, SES and SBA and the MSR cost function for TSB.

The next method, which will be used mainly as another benchmark, is Simple Exponential Smoothing (SES). As Croston was aimed at outperforming SES for intermittent demand, it can be expected that the performance of SES may be less than Croston in terms of forecasting, but that SES can outperform Croston in terms of service level, which is why this method is included. Another reason is that SES is commonly used in practice (Gardner Jr, 2006, Rostami-Tabar et al., 2013). The weighted average predictions from SES decrease over time and are smoothed by the smoothing parameter a, which is constructed by the cost function. This smoothing parameter is usually set somewhere between 0.1 and 0.3 in a setting with intermittent demand (Syntetos and Boylan, 2005). The formula for SES is

$$\hat{y}_t = ay_t + (1-a)\hat{y}_{t-1}. \tag{2.2}$$

As Syntetos and Boylan (2005) showed that Croston's method is biased, their method (SBA) is also included. In their approximation they introduced a smoothing parameter a that attempts to reduce the bias and it is aimed at smoothing the inter-demand interval p_t. SBA's formula is

$$\hat{y}_t = (1-\frac{a}{2})\frac{\hat{z}_t}{\hat{p}_t}. \tag{2.3}$$

Similarly to Croston, the first observation provides the initial values of z_t and p_t and the smoothing parameter a is set to 0.1.

The next method is TSB, introduced by Teunter et al. (2011). Their method criticized Croston's method on it's poor performance on obsolescence, which is mainly caused by the relatively slow updating, happening only when sales occur. In intermittent demand with obsolescence, this may cause the obsolescence to be caught onto very late. This is why TSB modifies Croston by replacing the inter-demand interval p_t with the d_t, the demand probability. d_t is 1 when demand does occur and otherwise it is 0. The forecast for TSB is again done through SES. The predictions by TSB are given by the formula

$$\hat{y}_t = \hat{d}_t\hat{z}_t. \tag{2.4}$$

As Willemain et al. (2004) showed empirically that their method outperformed SES and Croston, their bootstrapping method is also included in the comparison. Their bootstrapping method consists of seven steps, starting with step 1, estimating the transition probabilities for the two-state Markov model from historical demand. The next step (2) is to use the Markov model to generate zero and nonzero values over the specified forecast horizon, based on the last observed demand. The following step (3) is to replace the nonzero demands with a random demand value, with replacement, from the already known set of nonzero demands. Because this causes the resulting values to only show the same values that are already present in the data set, their next step (4) is to jitter the values of the nonzero demands, meaning a value close to the randomly selected demand value is selected to simulate a more natural variation in demand sizes. After the jittering process, the predicted values are summed over the forecast horizon (step 5), resulting in one prediction of the lead-time demand (LTD). Steps 2-5 are repeated many times to result in many LTD values, which are then sorted and the resulting distribution of the LTD can then be used.

As the M4 competition and the interpretation of the results by Makridakis et al. (2018, 2020a) showed that machine learning could provide the field of forecasting with interesting new methods, a simple neural network method following the methodology of Spiliotis et al. (2020) is constructed to evaluate the performance of a simple machine learning method and to compare it against the more well known methods. Spiliotis et al. (2020) construct a Multi-Layer Perceptron (MLP), which can also be referred to as a single hidden layer neural network. In order to train the model, the standard approach by Smyl (2020) was used, where a rolling input and output window and constant size are adopted. This means that a set amount of data points are used to predict future data and then once a future data point has been predicted, this will be added to the set amount of data points used to predict the following data point, while the very first data point in the set is dropped to make space for the newer data point. The data have to be scaled when the MLP is applied because the MLP applies a nonlinear activation function. This is done because the nonlinear activation function may encounter computational problems otherwise and because this increases the learning speed (Zhang et al., 1998). The data is linearly transformed to be scaled between 0 and 1 according to

$$y_t^{,} = \frac{y_t - y_{min}}{y_{max} - y_{min}} \tag{2.5}$$

and this scaling is reversed after the predictions have been made to obtain the final predictions and evaluate the forecasting accuracy. To construct the MLP, the R-package RSNNS will be used (Bergmeir et al., 2012).

The last method used is based on the LightGBM algorithm and the code for constructing the predictive model is adapted from Kailex (2020), which was created as an entry to the M5 competition. The main reason for adding this model to the comparison is to see whether the introduction of intermittent data in the M5 competition results in approaches which are suitable for spare parts demand forecasting. Other reasons for including this specific method are that LightGBM was the base for many of the top methods introduced in the M5 competition and the availability of the code, as not all top methods have shared their code or made it available in R (Makridakis et al., 2020b). As with the MLP method described above, a rolling input and output window are used and lag variables are constructed which are later used for the forecasts. The model is trained according to a Poisson loss and the hyper parameters for the LightGbm algorithm were adapted from Kailex (2020). The model is then trained and the training iterations are evaluated based on the Root Mean Squared Error (RMSE) at every 400 iterations. The training stops once the optimal RMSE has been found. The next step is to predict, where the lag variables are used to predict one day ahead at a time. For more information on the inner workings of LightGBM and the R implementation, please refer to the documentation by Microsoft (2021).

2.3 Selected forecasting accuracy measures

To verify whether the predictions made by the different methods are accurate compared to the actual values, several forecasting accuracy measures are taken into account, as is done in most comparative papers according to (Pinçe et al., 2021). Pinçe et al. (2021) also find that the most common forecasting accuracy measures used in recent spare parts demand literature are the absolute accuracy measures. For this paper, the absolute accuracy measures Mean Squared Error (MSE_t) and Mean Absolute Scaled Error $(MASE_t)$ will be used, which are all functions of the forecast errors $e_s = Y_s - \hat{Y}_s$. MSE_t is defined as

$$MSE_t = \frac{1}{t}\sum\nolimits_{s=1}^{t} e_s^2 \tag{2.6}$$

and $MASE_t$ is defined as

$$MASE_t = \frac{1}{t}\sum\nolimits_{s=1}^{t} \frac{|e_s|}{\frac{1}{t-1}\sum_{i=2}^{t}|Y_i - Y_{i-1}|}. \tag{2.7}$$

Derived from the M5 forecasting accuracy competition and originally proposed by Hyndman and Koehler (2006), the Root Mean Squared Scaled Error $(RMSSE_t)$ will also be used (Makridakis et al., 2020b). The $RMSSE_t$ is defined as

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(y_t - \hat{y}_t)^2}{\frac{1}{n-1}\sum_{t-2}^{n}(y_t - y_{t-1})^2}} \tag{2.8}$$

where h is defined as the forecasting horizon (which is set to the length of the test data set), y_t and $\hat{y}_t$ the actual and the predicted values of the time series at point t in time respectively and n as the amount of observations in the training sample.

After all of the previous accuracy measures have been obtained, the relative accuracy measure *Percentage Better* will be used to determine which method performed better than which other method on which data set and on which accuracy measure (Pinçe et al., 2021). *Percentage Better* is defined as the amount of times a method outperforms another method divided by the total amount of times the methods were applied simultaneously.

2.4 Selected inventory control measures

As the field of spare parts demand forecasting is not only guided by the need for accurate forecasting but also by the goal of keeping inventory costs low and achieving high service levels, two inventory control measures were included. As Van Wingerden et al. (2014) found, comparing methods solely on forecasting performance may overshoot their practical relevance and that is why service levels and tradeoff curves are identified as practically relevant inventory control measures. Pinçe et al. (2021) also identify tradeoff curves and service levels as the most used inventory control measures in recent spare parts demand literature. As the service level measure, fill rate will be used. More specifically, the item fill rate and not the order fill rate, as done by Van Wingerden et al. (2014). For the tradeoff curves, the tradeoff between the achieved fill rate and the holding costs will be used as the measures. Holding costs are defined as 25 percent of the price per year (Durlinger and Paul, 2012). To measure the achieved fill rate, first an inventory policy has to be determined. For this paper, the approach by Van Wingerden et al. (2014) was adapted, where a base stock level R is determined by evaluating past demand. The Inventory Position (IP) is updated each period, where back ordering is allowed. The IP is defined as

$$\text{IP} = stock\,on\,hand + outstanding\,orders - back\,orders. \tag{2.9}$$

The Inventory Level (IL) could also be defined as

$$\text{IL} = stock\,on\,hand - back\,orders. \tag{2.10}$$

If IP falls below R, an amount is ordered such that IP is equal to R. Van Wingerden et al. (2014) also indicate that a minimum order quantity may be a factor, but for simplicity this is not included in the approach in this paper. An example calculation for the Part Fill Rate (PFR) is as follows: if the demand size is 10 and the IP is 5, the achieved PFR will then be 50% at that demand moment. This is averaged across all demand moments for each series to determine the achieved Fill Rate for each method.

To generate the tradeoff curves, first a target fill rate will be set. Then the corresponding base stock level R will be computed according to the forecast method. The fill rate targets used for this paper will be 75%, 80%, 85%, 90%, 95%, 99% and 99,9999%, where the final fill rate target is interpreted as 100%. Next, the fill rate performance of the method will be evaluated compared to the target fill rate and the holding costs associated with each fill rate. These steps will be repeated for every forecasting method to generate the tradeoff curves.

2.5 Demand classification and training and test splits

As demand may show different characteristics and these differences in characteristics may improve a method's performance, Boylan et al. (2008) provide a framework to classify demand in four distinct groups based on the mean inter-demand interval p and CV^2, which is the squared coefficient of variation of the demand sizes. These two indicate whether the demand of an item can be classified as either Erratic ($p < 1.32$ and $CV^2 >= 0.49$), Lumpy ($p >= 1.32$ and $CV^2 >= 0.49$), Smooth ($p < 1.32$ and $CV^2 < 0.49$) and Intermittent ($p >= 1.32$ and $CV^2 < 0.49$). Erratic demand items are further defined to have highly variable demand sizes, Intermittent demand items show very infrequent demand occurrences of mainly the same values, Lumpy demand items are intermittent demand items with highly variable demand sizes and Smooth demand has frequent demand with low demand size variability (Boylan et al., 2008).

The calculation of p for each individual item is

$$p = \frac{Total\ number\ of\ time\ periods}{Count\ of\ the\ non\ zero\ demands} \tag{2.11}$$

and the calculation of CV for each individual item is

$$CV = \frac{Standard\ deviation\ of\ the\ non\ zero\ demands}{Mean\ of\ the\ non\ zero\ demands} \tag{2.12}$$

after which CV^2 can be obtained by squaring the result of equation 2.13. After obtaining the results, each individual item can be labeled as either erratic, lumpy, smooth or intermittent. To investigate whether the data that will be used for this paper show the aforementioned characteristics, in the next chapter the data sets will be classified according to the classification scheme by Boylan et al. (2008).

To be able to apply the performance measures mentioned above, each data set is split into a training and a test data set. The first 70% of the original data set will be used as the training data set for the forecasting methods. The last 30% of the data set will be used to compare the predictions of the forecasting method with, which will be referred to as the test data set.

Chapter 3

Data description and classification

As the experimental design is now clear, this section will outline the main characteristics of the data sets to which each method will be applied. Also, the origin of the data and all of the changes made to them will be explained. As this paper uses both industrial and simulated data and their initial set-up and origin slightly differs, the sections are split up into industrial data sets and simulated data sets. In order to improve reproducibility, all of the data sets will be available (in both their unaltered and altered form) on the dedicated GitHub page by de Haan (2021).

3.1 Industrial data sets

Four industrial data sets will be used for this paper. The first data set contains sales of 3451 items from a manufacturing company based in the Netherlands. The data was gathered over a period of 150 weeks, starting at the first week of 2012 and ending on the 46th week of 2014. The data set contains prices, inventory costs (set to 20% of the product cost), lead time, demand frequency and demand size data, amongst other variables such as a minimum order quantity and fixed order costs. The second data set is from the British Royal Air Force, containing sales of 5000 aircraft spare parts over the course of seven years (1996-2002). Again, prices, demand size and frequency and lead time are all available, but inventory costs are not determined. This data set was previously used by Teunter and Duncan (2009). The third industrial data set is from the automotive industry, with sales on 3000 items over the course of two years. This data set is the same data set used by Syntetos and Boylan (2005) and does not contain price or lead time information. The fourth and final industrial data set is sales data on 14523 spare parts for an oil refinery. The data sets contains monthly sales for each item for the period from January 1997 up to August 2001, thus spanning 56 months. This is the same data set that was used by Porras and Dekker (2008). Prices and lead times are available and this data set also provides insight into the current stock policy, with the minimum and maximum stock amounts shown for each item and the data set also indicates whether the system in which an item is installed in is considered low, medium or highly critical to operations. In order to distinguish the data sets, they will be named

Table 3.1 Descriptive statistics for the MAN, BRAF, AUTO and OIL data sets.

Data	Monthly item sales				Product price			
	min	mean	max	SD	min	mean	max	SD
MAN	0	10,39	4599,65	92,05	€ 0,03	€ 35,96	€ 2669,70	€ 101,81
BRAF	0,04	1,44	65,08	3,63	£ 0,001	£ 102,32	£ 9131,99	£ 373,29
AUTO	0,54	4,45	129,17	7,57	€ 26,82*	€ 778,54*	€ 6396,01*	€ 1126,58*
OIL	-182	1,00	180,46	6,13	€ 0,01	€ 450,34	€ 82562,59	€ 1453,91

*Added by using the *RPS* and *RMS* calculations described in section 3.1.

MAN, BRAF, AUTO and OIL respectively for the purposes of this paper. Table 3.1 gives an insight into the differences and basic statistics of the industrial data sets. As the MAN data set shows weekly sales, a 4-week period was used as the monthly sales number. As side notes: the negative minimum of the monthly item sales for the OIL data set are caused by returns and the minimum price of £ 0,001 for the BRAF data set is explained by minimum buying quantities, where a number of the same product are sold at a low price, thus causing the actual price to be below one cent.

In order to add pricing data to the AUTO data set, the relationship between pricing and monthly order frequency in the other data sets was examined. This was done in order to be able to calculate the inventory control performance of the methods on this data set. The way the relationship was examined is by looking at the ratio *RPS* (Ratio Price and Sales) between the average product price and monthly sales for each data set. The calculation of the *RPS* for each data set can be denoted as

$$RPS = \frac{Average\ item\ price}{Average\ monthly\ item\ sales}. \tag{3.1}$$

Although this method does not consider that the automotive industry spare parts might be differently priced relative to the other industries, this at least provides the data set with a price to order frequency ratio. The ratios of the MAN, BRAF and OIL data sets and the set ratio for the AUTO data set can be found in table 3.2. The *RPS* for the AUTO data set was set to the average *RPS*, which is 174.952. By multiplying the monthly sales by the *RPS*, the average product price was calculated for the AUTO data set. The mean product price for the AUTO data set turned out very high, which is likely due to the fact that while the average monthly item sales are relatively low, the data set also contains a portion of higher frequency items, which greatly affect the average product price. To determine the individual product prices and thus also establish a standard deviation for the AUTO data set, simply multiplying the individual monthly item sales with the *RPS* would result in high frequency items being extremely expensive. From examining the other data sets, it is clear that high frequency items are generally lower priced

Table 3.2 Ratio between price and monthly sales for the MAN, BRAF, OIL and AUTO data sets.

Data set	*RPS*
MAN	3.461
BRAF	71.056
OIL	450.340
AUTO	(set to) 174.952

items. In order to correctly attribute the individual item prices, this inverse relationship between product price and frequency needs to be respected. This is done by calculating the ratio *RMS* (Ratio Monthly Sales) between the monthly item sales for each individual item and the mean monthly item sales of the data set and dividing the average product price by the resulting number. The *RMS* can be calculated for each item through the formula

$$RMS = \frac{Average\ monthly\ individual\ item\ sales}{Average\ monthly\ item\ sales} \tag{3.2}$$

after which the individual item price can be calculated through the formula

$$\text{Individual item price} = \frac{Average\ item\ price}{RMS} \tag{3.3}$$

So in the example of the AUTO data set, if a product is sold an average of two times every month, the RMS can be calculated as $RMS = \frac{2}{4{,}45} \approx 0.449$. The average price for the AUTO data set is € 778,54, so dividing € 778,54 by 0.449 would result in an individual product price of € 1733,94. This makes sense, as the product is sold more than twice as little as the average product in the data set, which indicates that the product is likely more expensive than the average product in the data set. Using this methodology, the AUTO data set was filled with individual prices for each individual item. Finally, the standard deviation was derived from these prices.

3.2 Simulated data sets

In order to assess whether the type of data influences the forecasting accuracy or inventory control performance of the forecasting methods, simulated data sets were also created. The creation of simulated data allows for the addition of more specific characteristics in the data, to test whether these characteristics influence the performance. In the case of spare parts demand forecasting, these characteristics include the presence of intermittent, lumpy, smooth or erratic demand. To make the four simulated data sets used in this paper, the R package ’tsintermittent’ was used. To create a data set with this package, there are several required input arguments. Firstly, the number of time series needs to be determined. The average amount of time series in the industrial data sets was 6493,5, so for the simulated data sets n will be set to 6500 so that the simulated data sets resemble the size of the industrial data sets. Next, the number of observations

Table 3.3 Settings for the simulated data sets.

Data set	Intended demand pattern	CV^2	p	Monthly demand		Product price*	
				mean	SD	mean	SD
SIM1	Erratic	0.75	1.00	10,01	1,12	€ 1751,27	€ 202,28
SIM2	Lumpy	0.80	1.50	6,66	1.12	€ 1165,18	€ 209,51
SIM3	Smooth	0.30	1.05	9,50	0,74	€ 1662,04	€ 130,97
SIM4	Intermittent	0.25	1.45	6,90	0,81	€ 1207,17	€ 148,38

*Added by using the *RPS* and *RMS* calculations described in section 3.1.

for each series has to be set, which translates to the amount of periods in the industrial data sets. The data will be set up as if monthly sales data and the amount of periods will be set to 60 months (five years).

Now that the size of the data sets is clear, the average demand size and the squared coefficient of variation CV^2 and average interval of the non-zero demands p in the data sets need to be chosen. As these parameters are used to classify the data sets in the next section, these are set as such so that the four simulated data sets will be classified in each of the four distinct categories, as described in section 2.5. The average demand size is set to the arbitrary number of 10 for all of the simulated data sets. The data sets are labeled as SIM1, SIM2, SIM3 and SIM4 and their settings can be found in table 3.3. The influence of the zero demands are visible, as the mean monthly demand is lower when the inter-demand interval p is higher than 1. As with the industrial data sets, the average *RPS* of 174.952 was used to determine the average product price, after which the individual product prices were determined through the process described in section 3.1. As mentioned in the documentation for the package and in the accompanying article by Petropoulos et al. (2014), the simulator assumes a Bernoulli distribution for the non-zero demand occurrences and a negative binomial distribution for the non-zero demands.

3.3 Classifying the data sets

In order to classify the industrial and simulated data sets, the classification scheme by Boylan et al. (2008) was used, which is based on the demand-based classification by Syntetos et al. (2005). They classify each item based on the squared coefficient of variation CV^2 and average interval of the non-zero demands p into erratic, lumpy, smooth or intermittent items. A visual representation containing the cut-off points of 0.49 and 1.32 respectively can be seen in figure 3.1. After calculating the CV^2 and p for each individual item (the AUTO data set already had this information available), each item was classified.

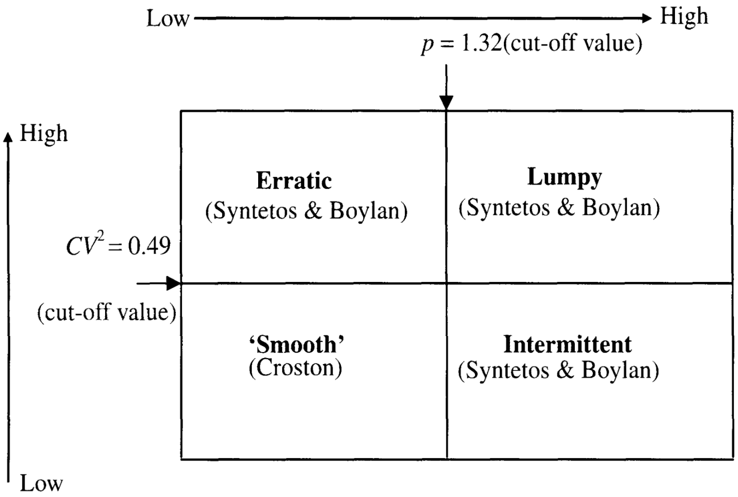


Fig. 3.1 Demand-based categorization scheme for forecasting by Boylan et al. (2008).

Table 3.4 Data set classifications.

Data set	CV^2	p	Erratic items	Lumpy items	Smooth items	Intermittent items
MAN	0.71	40.99	23	879	1	1038
BRAF	0.63	11.14	0	2095	0	2905
AUTO	0.44	1.30	441	314	1271	974
OIL	0.56	25.75	0	4402	0	9475
SIM1	0.75	1.00	6198	0	302	0
SIM2	0.80	1.50	410	5614	25	451
SIM3	0.30	1.05	36	0	6464	0
SIM4	0.25	1.45	1	7	786	5706

The results from the classification can be seen in table 3.4. The inter-demand interval for the industrial data sets is significantly higher than those of the simulated data sets, which indicates a higher degree of intermittency in the demand occurrences. The industrial data sets also contain very little erratic or smooth items, except for the AUTO data set. The AUTO data set also has the largest spread across all of the categories, with most of the demand patterns being smooth, which is also reflected in the relatively low inter-demand interval of 1.30. A side note on the results for the OIL and MAN data is that the total amount of classified items does not match with the total amount of items in the data sets. This is due to the presence of items with zero demand for all of the periods mentioned in the data sets, which results in the inability to classify them. The RAF and AUTO data sets do not contain items with zero demand for every period.

For the simulated data sets, it can be seen that the intended demand patterns as mentioned in table 3.3 have been correctly simulated. SIM1 generally contains erratic items, SIM2 mostly has lumpy items, SIM3 contains smooth items and SIM4 intermittent items. The other demand patterns are also present in some of the simulated data sets, which happens due to the scattering nature of the simulation procedure which may result in items being classified differently than the mainly intended classification. Applying the different methods to each data set will show which methods work best on which types of data and if the methods are vulnerable to items with different demand patterns than the one they are intended to deal with. In the following chapter of this paper, the application and results of the methods will be presented.

Similarly to Willemain et al. (2004), although negative values (interpreted as returns) can be accommodated by the aforementioned methods, the negative values have been coerced to zeroes for this paper when applying the methods. Their reasoning for this coercion is that returns should in theory be driven to zero in the long term and the replacement of negative values with zeroes serves as a conservative alternative, because otherwise the returns would be regarded as a stock replenishment. Also, all methods require at the least two demand occurrences to calculate the forecasts. Therefore, 6847 items were dropped from the training and test data sets for the OIL data set and 2059 items were dropped for the MAN data set. The other industrial data sets did not need items removed. Similarly, although the lead times are available for some of the industrial data sets, the lead time is set to 1 for the purpose of this paper. This is done to simplify the forecasting accuracy and inventory control performance assessment of each method.

Chapter 4

Results

4.1 Example of each forecasting method

In order to show how each method works, this section will show step-by-step how each method processes and learns from the input data and how the forecasts are then produced. This section will also show how any available parameters that apply to each method were chosen or adapted. To be able to see how the methods differ in the way they process the same input data, a random series from the SIM4 data set serves as the input data for this entire section. The series used is ts.3 and the characteristics for ts.3 can be seen in Table 4.1 and the first twenty data points can be seen in Table 4.2. As ts.3 is from one of the simulated data sets, it contains demand data for 60 periods. For all of the methods, the data is split into a training and test data set with a 70/30 ratio. This means that the forecasting horizon will be set to the length of the test split, which now contains 18 periods.

4.1.1 Croston, SES, SBA and TSB

For Croston, the input data starts with the 42 values in the training data. Croston requires the input of the forecast horizon, smoothing parameters, the initial values for demand and interval size and which cost function should be used for the optimization. The forecast horizon is set to 1 (the lead time), making the forecasts lead-time demand (LTD) forecasts, while the smoothing parameters are optimised by the cost function, which is the Mean Absolute Rate (MAR), as this was found to be the optimal cost function for Croston by Kourentzes (2014). This means that the

Table 4.1 The characteristics of ts.3 from the SIM4 data set.

Name	CV^2	p	Demand				Price	Classification
			Minimum	Median	Maximum	Average		
ts.3	0.22	1.58	0	6.5	23	10,29	€ 809,51	Intermittent*

* CV^2 is less than 0.49 and p is more than 1.32.

Table 4.2 The first twenty ts.3 data points for the first twenty periods.

ts.3	Period																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Demand	0	8	7	7	0	0	18	6	11	6	0	4	9	23	0	13	0	0	0	0

smoothing parameters are set to such values that the lowest possible MAR is achieved. The first non-zero demand was used as the initial demand and the initial interval is set to the first interval. As Croston, SES, SBA and TSB are similar in their approach and simplicity, the process and settings are exactly the same, with the only difference that for TSB, the Mean Squared Rate (MSR) was used as the cost function. After a prediction has been made by the methods, the predicted demand value is added to the training data and the process starts again, until all 18 values have been forecast.

4.1.2 Willemain

For the Willemain bootstrapping method by Willemain et al. (2004), the process starts with converting the training data to zeroes and ones, the periods with zero demand are a zero and the periods with a positive demand are a 1. Then the transition probability matrix can be determined. This contains the conditional probability of the series whether the next period will have a demand or not depending on the previous period. For ts.3 the transition probabilities are as follows: if the previous demand was a zero, then the next demand has a 35,7% chance to be another zero demand period, and a 64,3% chance to be a positive demand. If the previous demand was a positive demand, then the next demand has a 29,6% chance to be a period without demand and a 70,4% chance to be another positive demand.

Once the transition matrix is determined, this can be used to generate a series of zeroes and non zeroes over the forecast horizon, which is set to the lead time. This means that the forecast from Willemain bootstrapping is again the lead-time demand (LTD). The last value in the training data set for ts.3 is 8, which is a positive demand. The method will then predict what will follow based on this positive demand 1000 times. From the transition matrix we can see that of these 1000 predictions, it is likely that around 30% will be a zero and around 70% will be a one (a positive demand). Next, all of the positive demands are replaced with a random positive demand value from the training data. From looking at Table 4.2 again, we can see what some of these values will be.

The next step is to jitter these values, which means that an 8 may now be transformed into for example a 9, a 7 or a 10. This is done to simulate a natural demand size variation. All of the predicted and jittered values are summed over the forecast horizon, resulting in one LTD value.

In this special case where the lead time is 1, the predicted and jittered values already are the LTD predictions. From these LTD values, the mean and standard deviation can be derived. For ts.3, the resulting mean and standard deviation from applying Willemain bootstrapping for the first forecast are 8.34 and 7.32 respectively. Note that the results may slightly differ each application due to the randomness of the jittering process, unless the amount of predicted LTD values is very large. After the forecast is obtained, the predicted value is added to the training data, classified as either a demand or zero demand again and the process starts anew.

4.1.3 Multi-Layer Perceptron (MLP)

For the neural network method, first the data is normalized to a scale of 0-1. This will be reverted after the predictions are made to end up with the final predictions. After the normalization, a rolling window is applied to the training data. 5 periods are taken for the input window, as this number is sufficient for the machine learning methods to learn the underlying dependencies, but small enough to provide the methods with enough data for the industrial data sets where series are short. If the window would be larger, it may mean that too few data points are available to learn from and that the methods are under trained. This means that the first 5 periods in the training data are saved as input data, with the 6th period as the first output data. The output window is not only 1 because it is equal to the lead time, but also in order to compare the MLP method with the statistical methods more fairly. Next, the window moves up one period, with periods 2-6 as the input data and period 7 as the output data. This process is repeated until no more input data is available, resulting in a new data set for each item where the first 5 columns represent the input data and the 6th column is the corresponding output. The first 10 rows of this new data set for ts.3 can be seen in Table 4.3. The final 5 rows have missing values, as no output is available for the input data. These rows are removed from the data set.

The next step is to train the neural network with the newly made data set. The neural network will learn the underlying relationships within the data and train itself based on those findings. The neural network offers a lot of options to adapt based on the data. For this paper, the maximum amount of iterations was found to be optimal at 200 and the size (amount of nodes) in the hidden layer at 6. The other hyperparameters used in the training of the model are set to the defaults, as suggested by Bergmeir et al. (2012). When the model is trained it can be used to predict the next demand value based on the last 5 periods. Therefore the last 5 periods from the training data set are used as input for the MLP and the next period is predicted. After this new prediction has been attained, the prediction is added to the 5 input periods and the first period in the input data is removed in order to be at 5 input periods again. These next period predictions are done until the end of the forecasting horizon is reached. As mentioned earlier, these predictions have to be denormalized again in order to represent demand values.

Table 4.3 Normalized rolling input and output data set for the MLP constructed for ts.3.

	Input 1	Input 2	Input 3	Input 4	Input 5	Output
1	0.00	0.3478261	0.3043478	0.3043478	0.00	0.00
2	0.3478261	0.3043478	0.3043478	0.00	0.00	0.7826087
3	0.3043478	0.3043478	0.00	0.00	0.7826087	0.2608696
4	0.3043478	0.00	0.00	0.7826087	0.2608696	0.4782609
5	0.00	0.00	0.7826087	0.2608696	0.4782609	0.2608696
6	0.00	0.7826087	0.2608696	0.4782609	0.2608696	0.00
7	0.7826087	0.2608696	0.4782609	0.2608696	0.00	0.1739130
8	0.2608696	0.4782609	0.2608696	0.00	0.1739130	0.3913043
9	0.4782609	0.2608696	0.00	0.1739130	0.3913043	1.00
10	0.2608696	0.00	0.1739130	0.3913043	1.00	0.00

4.1.4 LightGBM method

For the LightGBM method, the input data is the same as the MLP input data in Table 4.3. The algorithm learns from the first 5 columns and column 6 serves as the output corresponding to those input values. For the LightGBM algorithm, several hyper parameters need to be setup. For this paper, the hyper parameter setup by Kailex (2020) is used. The algorithm, once initiated, goes through training rounds and stops when the method attains the lowest RMSE. Similar to the MLP, once the model is trained, it can be used to predict the next period based on the previous 5 periods. The last 5 periods from the training data set are therefore again used as input and the next period predictions are again done until the end of the forecasting horizon is reached. Once the predictions have been denormalized, the 18 predicted demand values are obtained for ts.3.

4.2 Examples of the accuracy measures

When all methods are applied to ts.3, the result is 18 predictions corresponding to the length of the forecast horizon. The original 18 data points in the test set for ts.3 and the predictions made by the forecasting methods can be seen in Table 4.4.

4.2.1 Forecasting accuracy

In order to compare the methods' accuracy, the forecasting accuracy measures MSE, MASE and RMSSE are then applied. The results can be seen in Table 4.5. For all of these metrics a lower value represents a better performance. The best performing method is highlighted for each metric. For this specific item, the LightGBM method outperforms the others in terms of MSE and RMSSE and Croston has the lowest MASE.

Table 4.4 Test set data and predictions by each method for ts.3.

Data	Period																	
	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
ts.3	0	0	0	0	0	12	0	7	7	12	0	8	8	8	0	9	6	10
Croston	6.70	6.60	6.84	6.79	7.13	7.16	7.20	7.20	7.22	7.21	7.22	7.22	7.22	7.23	7.23	7.23	7.23	7.23
SES	7.40	7.40	7.40	7.40	7.37	7.40	7.37	7.37	7.37	7.37	7.37	7.37	7.37	7.37	7.37	7.38	7.38	7.38
SBA	6.48	6.34	6.20	6.09	6.04	6.03	6.31	6.59	6.74	6.82	6.86	6.94	7.01	7.01	7.01	7.00	6.98	6.96
TSB	7.56	7.64	7.62	7.66	7.68	7.72	7.73	7.74	7.74	7.73	7.73	7.73	7.73	7.73	7.72	7.71	7.70	7.70
Willemain	8.34	8.35	8.26	8.47	8.95	9.01	8.87	9.26	9.05	9.02	9.40	9.06	9.23	9.67	9.50	9.48	9.53	9.59
MLP Method	6.69	6.69	6.79	6.74	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71
LightGBM	4.56	7.42	4.25	5.93	3.85	5.84	3.33	3.91	3.56	4.26	4.11	6.94	6.78	5.94	4.24	6.36	2.79	3.97

Results rounded to two decimals.

Table 4.5 The forecasting accuracy of all methods on item ts.3 predictions.

Method	MSE	MASE	RMSSE
Croston	24.942	**0.567**	1.856
SES	27.331	0.588	1.943
SBA	26.134	0.579	1.900
TSB	28.791	0.595	1.994
Willemain	36.786	0.682	2.254
MLP Method	23.745	0.576	1.811
LightGBM	**20.535**	0.571	**1.684**

Results rounded to three decimals.

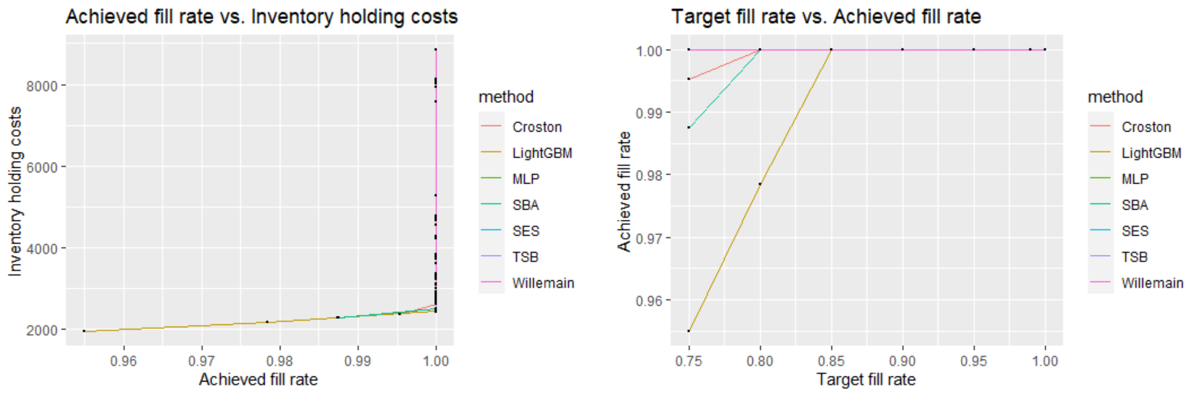


Fig. 4.1 Tradeoff curves for the inventory control measures on ts.3 predictions by Croston, SES, SBA and TSB

4.2.2 Inventory control performance

Subsequently, the inventory control performance of each method is assessed. First the base stock levels R for each target fill rate are determined according to the predictions made by each method. Next the holding costs and the achieved fill rates associated with those base stock levels can be calculated. Figure 4.1 shows the results for each target fill rate on the predictions made for ts.3. While these tradeoff curves are normally more gradual in their appearance when they are applied to a larger data set and then averaged, for a single item the results may vary. For this specific item, it can be seen that the achieved fill rate is very high compared to the target fill rate for all methods except for the LightGBM method and that the inventory costs required to obtain a 100% fill rate are to the lower side of the graph. This suggests that the test data for this item has a substantially lower demand than the training data, which results in the relatively high part fill rate. A more practical interpretation is that the methods overestimate the amount of demand that is expected.

4.3 Overall results

After all of the methods have been applied to the full data sets in the manner described above, the overall results are obtained. In order to also compare the methods on their ease of implementation in practice, the total and per item run time for each method was recorded. The results can be seen in Table 4.6. These run times and the ease of implementation in general will be further discussed in Chapter 5.

Table 4.6 Total run time and run time per item for each method (in R).

Method	Total run time	Average run time per item*
Croston	311 min	0,433 s
SES	309 min	0,430 s
SBA	309 min	0,430 s
TSB	48 min	0,067 s
Willemain	1198 min	1,669 s
MLP	175 min	0,243 s
LightGBM	240 min	0,334 s
Total	2590 min = 42.2 hours	

*Calculated over all 43068 items.

4.3.1 Forecasting accuracy

Similarly to the example shown previously on the ts.3 series, first each method is assessed based on the forecasting accuracy. The forecasting accuracy of each method on each data set can be seen in Table 4.7. To show how the methods rank comparatively for each data set, the Percentage Better results can be found in Table 4.8. This shows how often a method outperformed another method as a percentage of the total amount of comparisons.

The results show that there there was not a single method that outperformed the others on every data set based on forecasting accuracy. Every method except for the LightGBM method scores best on some measure on some data set. Surprisingly, the LightGBM method did not perform as well as it did in the example and it also performed the worst overall. What also stands out is that the MLP method performed well on the simulated data sets, which may be accredited to the way the data was pre-processed and the application of the method to aggregated data. This also explains why the MLP method does not perform well on the industrial data, since the data varies much more in terms of demand size and intermittency. This inhibits the learning for the method and makes the results more generalised, resulting in poorer results. For the industrial data sets, the parametric methods perform the best, with Willemain performing best on one metric on the MAN data set. This is again mainly caused by the approach before applying the method. For these methods, the approach was to input a single series and expand on this series with the predicted demands. This approach seems to work particularly well for industrial data. The performance of Willemain on the MAN data set may be due to the extremely high average time between demands that the data set showed. This makes all of the methods perform poorly on forecasting accuracy as they all overestimate the future demand occurrences even though the forecast demand amounts may be more accurate. This is again a problem that arises from the way the data is processed before applying the methods.

Looking into the Percentage Better results based on the forecasting accuracy we can see from the last column in Table 4.8, the average Percentage Better score, how each method performs

Table 4.7 The forecasting accuracy of all methods on each data set.

Method	Measure	Data set							
		SIM1	SIM2	SIM3	SIM4	MAN	BRAF	AUTO	OIL
Croston	MSE	79.257	79.277	**34.633**	40.698	13146.96	207.470	86.095	**860.847**
	MASE	0.670	1.014	0.486	0.780	2.372	2.123	0.783	8.107
	RMSSE	2.727	3.353	**1.876**	2.417	5.464	3.423	1.731	14.747
SES	MSE	79.273	79.852	34.919	40.907	12975.757	**198.206**	85.181	899.428
	MASE	0.670	1.020	0.487	0.783	2.165	1.881	0.787	7.865
	RMSSE	2.727	3.366	1.883	2.423	**5.243**	**3.296**	1.734	14.693
SBA	MSE	78.960	79.350	34.723	40.876	13126.803	202.178	**84.768**	871.933
	MASE	**0.659**	**0.997**	**0.483**	0.780	2.111	1.8404	**0.777**	7.829
	RMSSE	2.718	3.349	1.877	2.421	5.370	3.345	**1.722**	14.702
TSB	MSE	79.763	79.936	35.011	41.057	13088.866	199.555	86.031	929.241
	MASE	0.672	1.023	0.488	0.783	**1.948**	**1.766**	0.790	7.802
	RMSSE	2.735	3.369	1.885	2.427	5.247	3.301	1.739	**14.685**
Willemain	MSE	**78.226**	80.300	35.135	42.033	**12955.828**	212.609	84.851	901.377
	MASE	0.690	1.076	0.500	0.796	2.689	2.654	0.945	**7.773**
	RMSSE	2.721	3.395	1.893	2.457	5.390	3.579	1.921	14.729
MLP	MSE	78.569	**78.686**	35.336	**39.996**	23009.153	201.105	86.010	994.130
	MASE	0.681	1.060	0.498	**0.779**	49.619	2.304	0.860	21.988
	RMSSE	**2.716**	**3.348**	1.892	**2.396**	19.389	3.352	1.796	19.791
LightGBM	MSE	96.932	98.591	42.152	49.867	47578.854	239.699	125.430	1211.790
	MASE	0.747	1.132	0.537	0.847	75.672	2.301	0.965	22.226
	RMSSE	3.017	3.743	2.069	2.667	32.012	3.743	2.091	21.948

Results rounded to three decimals. The best accuracy is highlighted for each data set and measure.

compared to the rest. The percentages are calculated column-wise, where each method is measured against every other method based on the achieved forecasting accuracy in Table 4.7. So the first value in the table (66.67%), shows that Croston was better in 66.67% of the cases when compared to each of the other methods when they were applied to the SIM1 data and their forecasting accuracy was measured with MSE. In the last column, the percentage better scores are averaged for each method and accuracy measure across all data sets. These results show that SBA is on average the best performing method relative to the methods it is compared against on each of the metrics. Again, it is clear that the LightGBM method is the worst overall.

A revised approach for the pre-processing of the data.

As the results from both Table 4.7 and 4.8 indicate that the methods are struggling with the way the data is provided to them, a second application of each method was performed. In this revised approach, the training data is no longer updated with the values forecast by each method after each period. Instead, the real values from the test data sets are added to the training data sets. This is not possible in practice when one would like to predict more than one lead-time demand ahead, but this approach creates a setting where after each predicted period the training data is updated in order to use the most recent data for the forecasts. So although the initial

Table 4.8 Percentage better comparison on forecasting accuracy of all methods on each data set.

Method	Measure	Data set								Average
		SIM1	SIM2	SIM3	SIM4	MAN	BRAF	AUTO	OIL	
Croston	MSE	66.67%	83.33%	100%	83.33%	33.33%	33.33%	16.67%	100%	64.58%
	MASE	66.67%	83.33%	83.33%	66.67%	50%	50%	83.33%	33.33%	64.58%
	RMSSE	33.33%	66.67%	100%	83.33%	33.33%	33.33%	83.33%	33.33%	58.33%
SES	MSE	33.33%	50%	66.67%	50%	83.33%	100%	66.67%	66.67%	64.58%
	MASE	66.67%	66.67%	66.67%	33.33%	66.67%	66.67%	66.67%	50%	60.42%
	RMSSE	28.57%	50%	66.67%	50%	100%	100%	66.67%	83.33%	68.16%
SBA	MSE	66.67%	66.67%	83.33%	66.67%	50%	50%	100%	83.33%	**70.83%**
	MASE	100%	100%	100%	66.67%	83.33%	66.67%	100%	66.67%	**85.42%**
	RMSSE	83.33%	83.33%	83.33%	83.33%	66.67%	83.33%	100%	66.67%	**81.25%**
TSB	MSE	16.67%	33.33%	50%	33.33%	66.67%	83.33%	33.33%	33.33%	43.75%
	MASE	50%	50%	50%	33.33%	100%	100%	50%	83.33%	64.58%
	RMSSE	16.67%	33.33%	50%	33.33%	83.33%	83.33%	50%	100%	56.25%
Willemain	MSE	100%	16.67%	33.33%	16.67%	100%	16.67%	83.33%	50%	54.17%
	MASE	16.67%	16.67%	16.67%	16.67%	33.33%	0%	16.67%	100%	27.09%
	RMSSE	66.67%	16.67%	16.67%	16.67%	33.33%	16.67%	16.67%	50%	29.17%
MLP Method	MSE	83.33%	100%	16.67%	100%	16.67%	66.67%	33.33%	16.67%	54.17%
	MASE	33.33%	33.33%	33.33%	100%	16.67%	16.67%	33.33%	16.67%	35.42%
	RMSSE	100%	100%	33.33%	100%	16.67%	50%	33.33%	16.67%	54.17%
LightGBM	MSE	0%	0%	0%	0%	0%	0%	0%	0%	0%
	MASE	0%	0%	0%	0%	0%	33.33%	0%	0%	4.17%
	RMSSE	0%	0%	0%	0%	0%	0%	0%	0%	0%

Results rounded to two decimals. The highest average Percentage Better for each method is highlighted.

approach has a more straightforward application in practice, this revised approach provides the methods with the actual data. As lead times are not always equal to the 1 that it was set to in this paper in practice, this revised approach is more straightforward when lead times are longer.

The results from the second application of each method can be found in Tables 4.9 and 4.10. These revised results are in italics when the result is an improvement compared to the initial application and again the best performances are highlighted. Although again each method except for the LightGBM method performed best on some metric and data set, the SBA method has the best performance in by far the most instances. This is not only the case for the simulated data sets but also for the industrial data sets. Also, not every method shows improved results with the revised approach, although every method has an improved result on some metric on some data set. This suggests that any overestimating that happened in the initial application may have caused the results to be better than they should have been and that this approach more truthfully measures the method's performance on a lead-time demand prediction, but that that is not the case for every series. The over-estimations that happened earlier for the Willemain, MLP and LightGBM methods are now also less extreme, which can be seen by the lesser MSE values, which indicates closer estimates on average. The method that showed the least improvements is the TSB method, for which the earlier over-estimations seemed to have improved the method's accuracy, indicating that TSB may be constantly underestimating the actual demand. This will be checked later when the trade-off curves are examined, as TSB should then show signs of underestimating the required stock. The overall implications and findings are further discussed in the next section.

4.3.2 Inventory performance and overall assessment

Next, the inventory performance is assessed and is related to the performance of the methods based on the results from the forecasting accuracy section. The trade-off curves are set up separately for each data set, as seen in Figures 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8 and 4.9. For every data set, a trade-off curve is made to assess the trade-off between the fill rate that was achieved by the method and the associated holding costs, labeled with an (a). Achieving a higher fill rate always leads to higher holding costs, with the figures maxing out at the 99.9999% fill rate level, which is the approximation of a 100% fill rate for this paper and where the achieved fill rate cannot realistically increase any further without extreme increases in holding costs. The other trade-off curve shows the trade-off between the achieved fill rate and the target fill rate that had been set, labeled with a (b). These curves will show a gradual increase of the achieved fill rate until the target fill rate reaches the 99.9999% level again. The achieved fill rate will not always reach the 100% approximation in every graph, as high demand occurrences prohibit this. This is the case for the MAN, BRAF and OIL data sets. The trade-off curves are analysed separately in

Table 4.9 The forecasting accuracy of all methods on each data set with the revised pre-processing.

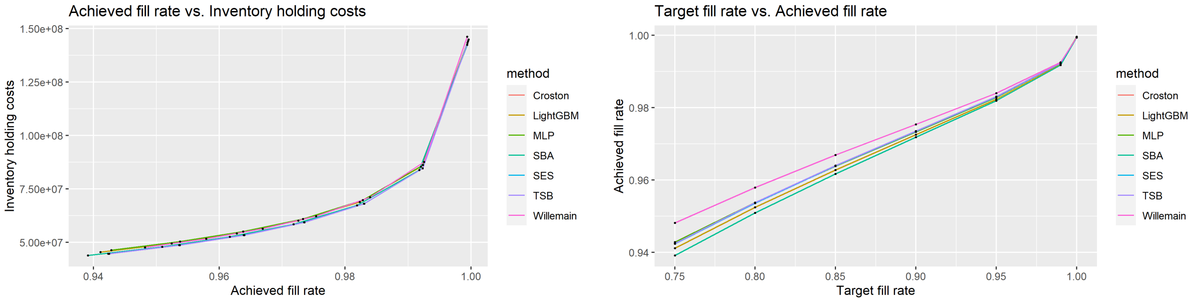
Method	Measure	Data set							
		SIM1	SIM2	SIM3	SIM4	MAN	BRAF	AUTO	OIL
Croston	MSE	*79.202*	*79.138*	34.730	*40.519*	*12940.104*	***199.690***	86.344	*708.262*
	MASE	0.673	1.027	0.487	0.780	2.499	*2.080*	0.788	8.331
	RMSSE	*2.722*	*3.346*	1.878	*2.412*	*5.329*	*3.300*	*1.721*	*12.515*
SES	MSE	*79.212*	*79.806*	*34.916*	40.938	13327.738	201.190	*84.445*	*701.551*
	MASE	0.673	1.028	0.488	0.783	2.335	1.996	*0.780*	8.283
	RMSSE	*2.722*	*3.358*	1.883	2.424	***5.215***	***3.283***	***1.710***	*12.471*
SBA	MSE	*78.623*	*78.834*	***34.620***	*40.435*	***12921.667***	*199.807*	***83.089***	***690.445***
	MASE	**0.664**	**1.012**	**0.484**	*0.778*	2.439	2.001	**0.777**	8.013
	RMSSE	***2.712***	*3.337*	***1.874***	*2.409*	*5.304*	*3.289*	***1.710***	***12.429***
TSB	MSE	*79.759*	79.966	35.091	41.149	13293.605	200.945	90.061	*706.219*
	MASE	0.675	1.032	0.490	0.785	**2.321**	**1.955**	0.792	8.013
	RMSSE	*2.731*	*3.363*	1.887	2.430	*5.223*	*3.287*	*1.728*	*12.482*
Willemain	MSE	**78.231**	*78.746*	*34.960*	*40.752*	13288.562	*200.354*	*84.622*	*695.419*
	MASE	0.690	*1.043*	*0.498*	*0.783*	*2.539*	*2.309*	*0.906*	***7.042***
	RMSSE	2.721	3.353	*1.889*	2.420	*5.321*	*3.381*	*1.874*	*12.634*
MLP	MSE	*78.500*	***77.511***	*34.736*	***39.747***	*14472.538*	202.271	*84.849*	*776.300*
	MASE	0.687	*1.032*	*0.491*	***0.776***	*20.673*	2.340	*0.836*	22.465
	RMSSE	2.716	***3.319***	*1.879*	***2.389***	*10.585*	3.369	*1.756*	*18.249*
LightGBM	MSE	98.352	*97.244*	43.316	*49.567*	*14672.193*	*207.561*	*115.453*	*811.872*
	MASE	0.758	1.141	0.547	0.849	*29.394*	2.354	*0.940*	*21.675*
	RMSSE	3.040	*3.720*	2.095	*2.660*	*14.727*	*3.413*	*2.070*	*18.899*

Results rounded to three decimals. The best accuracy is highlighted for each data set and measure.
The result is in italics if it was an improvement over the first application.

Table 4.10 Percentage better comparison on forecasting accuracy of all methods on each data set with the revised pre-processing.

Method	Measure	Data set								Average
		SIM1	SIM2	SIM3	SIM4	MAN	BRAF	AUTO	OIL	
Croston	MSE	50%	50%	83.33%	83.33%	66.67%	100%	33.33%	33.33%	62.50%
	MASE	66.67%	83.33%	83.33%	66.67%	50%	50%	66.67%	33.33%	63.49%
	RMSSE	33.33%	66.67%	83.33%	66.67%	33.33%	50%	66.67%	50%	56.24%
SES	MSE	33.33%	33.33%	50%	33.33%	33.33%	33.33%	83.33%	66.67%	45.83%
	MASE	66.67%	66.67%	66.67%	33.33%	83.33%	83.33%	83.33%	50%	*66.66%*
	RMSSE	33.33%	33.33%	50%	33.33%	100%	100%	100%	83.33%	66.67%
SBA	MSE	66.67%	66.67%	100%	83.33%	100%	83.33%	100%	100%	***87.50%***
	MASE	100%	100%	100%	83.33%	66.67%	66.67%	100%	66.67%	**85.42%**
	RMSSE	100%	83.33%	100%	83.33%	66.67%	66.67%	100%	100%	***87.50%***
TSB	MSE	16.67%	16.67%	16.67%	16.67%	50%	50%	16.67%	50%	29.17%
	MASE	50%	33.33%	50%	16.67%	100%	100%	50%	66.67%	58.33%
	RMSSE	16.67%	16.67%	33.33%	16.67%	83.33%	83.33%	50%	66.67%	45.83%
Willemain	MSE	100%	83.33%	33.33%	50%	66.67%	66.67%	66.67%	83.33%	*68.75%*
	MASE	16.67%	16.67%	16.67%	33.33%	33.33%	33.33%	16.67%	100%	*33.33%*
	RMSSE	66.67%	50%	16.67%	50%	50%	16.67%	16.67%	33.33%	*35.42%*
MLP Method	MSE	83.33%	100%	66.67%	100%	16.67%	16.67%	50%	16.67%	*56.25%*
	MASE	33.33%	33.33%	33.33%	100%	16.67%	16.67%	33.33%	0%	33.33%
	RMSSE	83.33%	100%	66.67%	100%	16.67%	33.33%	33.33%	16.67%	*56.25%*
LightGBM	MSE	0%	0%	0%	0%	0%	0%	0%	0%	0%
	MASE	0%	0%	0%	0%	0%	0%	0%	16.67%	2.08%
	RMSSE	0%	0%	0%	0%	0%	0%	0%	0%	0%

Results rounded to two decimals. The highest average Percentage Better for each method is highlighted.
The average Percentage Better is in italics if it was an improvement over the first application.



(a)

(b)

Fig. 4.2 Tradeoff curves for the inventory control measures on SIM1 predictions by every method.

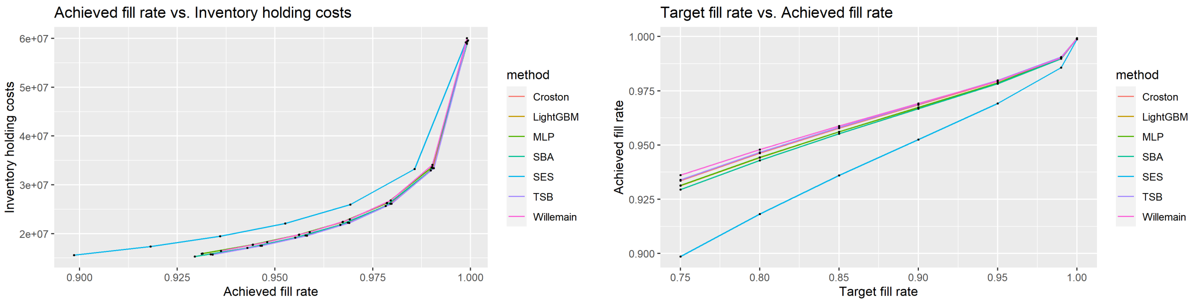
order to assess the performance of the used forecasting methods in the context of the inherent characteristics of each data set. When the results are related to the forecasting accuracy, the results from the revised approach are used in the assessment.

SIM1 inventory performance

The SIM1 data set was classified as a data set with erratic demand, meaning low intermittency but a highly variable demand size. The low intermittency means that the methods all achieve very similar results on the inventory holding costs assessment, as the predictions made by each methods are similar and stable. Therefore, no clearly superior method can be deducted visually from Figure 4.2a. However, when inspecting Figure 4.2b, there is a slight difference in the achieved fill rate of each method. The Willemain method achieves the best fill rate on every target fill rate leading up to the highest level, with the rest of the methods very close together. SBA performs worst on this assessment. The implication of the difference in achieved fill rate is that a slightly higher achieved fill rate does not equate to a significant increase in holding costs, as was seen from Figure 4.2a. This means that Willemain is the superior method for this application and this data set when these measures are used. This also shows how the forecasting accuracy assessment should not be the only metric, as SBA performed best on forecasting the SIM1 data set based on the MASE and RMSSE, but performs worst on the inventory control assessment. Willemain also performed best on forecasting the SIM1 data set based on MSE, which is now further supported based on the inventory control measures.

SIM2 inventory performance

Figure 4.3a shows a clearer separation. The blue line displaying the SES method is located above all of the other methods in the graph, indicating that for every achieved fill rate, higher holding costs are associated. Figure 4.3b also shows the lacking performance of SES, as the achieved fill rate is the lowest for every target fill rate. The rest of the methods are close together, similarly to the SIM1 trade-off curves, with Willemain being the superior method again by a small margin.



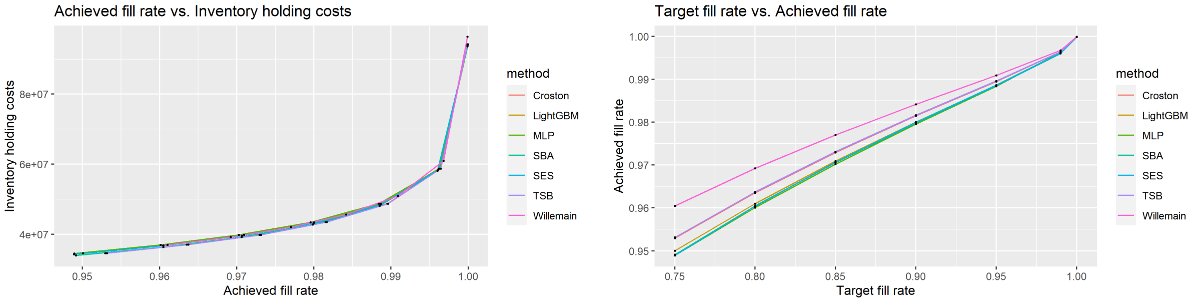
(a) (b)

Fig. 4.3 Tradeoff curves for the inventory control measures on SIM2 predictions by every method.

Surprisingly, the lacking performance of SES cannot be related to it's forecasting accuracy performance on the SIM2 data set, as it's performance there did not show lacking results. As the SIM2 data set was classified as mainly lumpy with some erratic and intermittent items, it mainly shows items with longer demand intervals. The addition of zeroes and the higher demand variability compared to the SIM1 data seems to have caused the SES method to underestimate the lead-time demand across all target fill rates, likely due to the averaging nature of the method. The superior forecasting accuracy portrayed by the SBA and MLP methods on the SIM2 data set does not appear to relate to a superior inventory control performance in this case.

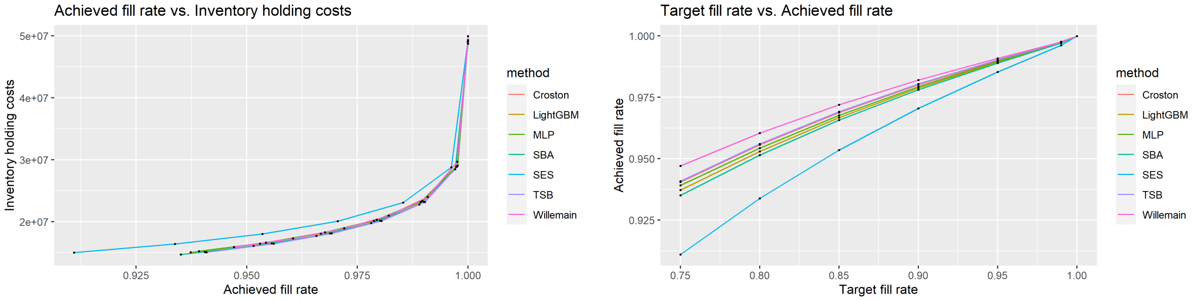
SIM3 inventory performance

The trade-off curves for the SIM3 data set in Figure 4.4 show similar results to the trade-off curves for the SIM1 data set. 4.4a shows each method relatively close in terms of the trade-off between holding costs and achieved fill rate, and 4.4b shows Willemain being the superior method in terms of achieving a higher fill rate. This time, the TSB and Croston methods are second and third and the rest of the methods are clumped together slightly lower than those two. Overall, especially Willemain, but also TSB and Croston are slightly better in achieving a slightly better fill rate with similar inventory holding costs. The SIM3 data set is classified as smooth, which means that demand variability is minimal and intermittency is also low. The similarity with the SIM1 results seems to suggests that the variability of demand size influences the results to a lesser extent than the variability of the average inter-demand interval, which was fairly similar in the SIM1 and SIM3 data sets (1.00 and 1.05 respectively), while the average variation in demand size did differ (0.75 and 0.30 respectively). Also, the absolute superiority of SBA on forecasting accuracy on the SIM3 data set predictions does again not relate to good inventory control performance, further suggesting that the two measures are separately important to anyone deciding which method best suits their application.



(a) (b)

Fig. 4.4 Tradeoff curves for the inventory control measures on SIM3 predictions by every method.



(a) (b)

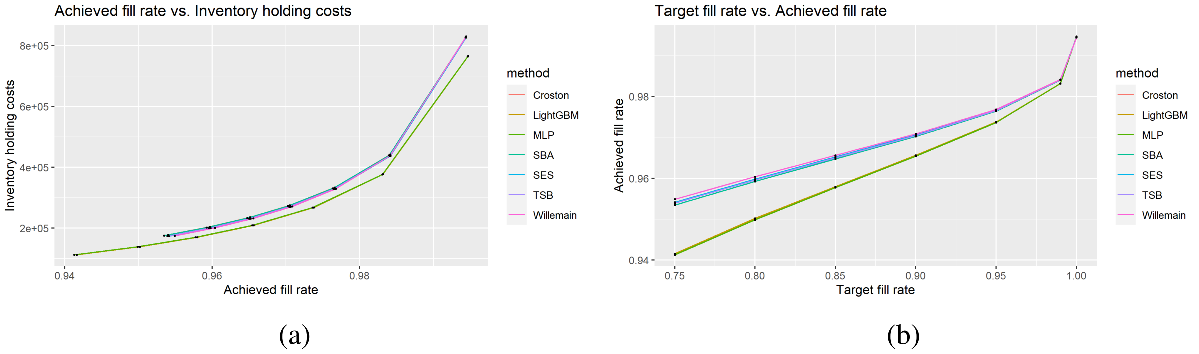
Fig. 4.5 Tradeoff curves for the inventory control measures on SIM4 predictions by every method.

SIM4 inventory performance

Similar to the similarity between the SIM1 and SIM3 results, the SIM4 results shown in Figure 4.5 resemble the results for the SIM2 data set. SES performs the poorest out of all of the methods, with higher inventory holding costs associated with each achieved fill rate and a lower achieved fill rate associated with each target fill rate. Also, the Willemain method is again the superior method in terms of both trade-offs. The similarity with the SIM2 results can this time be accredited to the other associated classification metric, the demand variability. Again the high demand variability seems to make the SES method perform poorly, with the rest of the methods clumped together and the Willemain method as the overall winner. For this data set, the MLP method was superior on all forecasting accuracy metrics, but this result does not relate to the inventory control performance in this case, as the method performs average relative to the other methods.

MAN inventory performance

Transitioning to the industrial data sets, the trade-off curves for the MAN data set in Figure 4.6 show a different result than the simulated data sets showed. In Figure 4.6a the MLP and LightGBM methods are to the lower left side of the graph compared to the other methods, indicating that they achieve lower inventory holding costs with the same achieved fill rate. This result is also reflected in Figure 4.6b, where the two methods achieve a lower fill rate for each



(a)

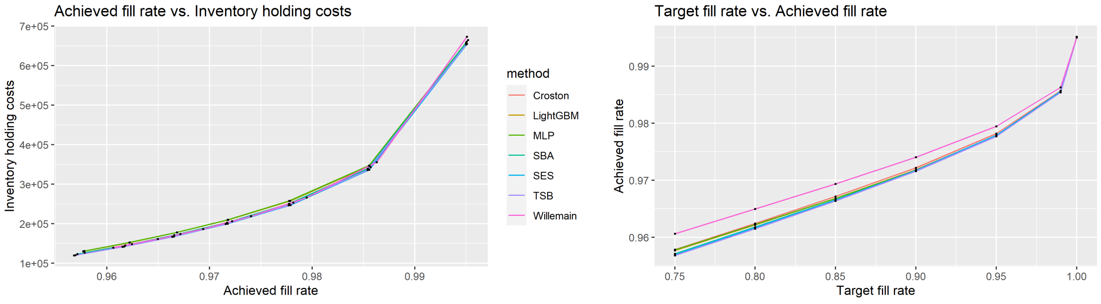
(b)

Fig. 4.6 Tradeoff curves for the inventory control measures on MAN predictions by every method.

target fill rate. This can be interpreted as a lesser extent to which the methods overestimate the required stock levels compared to the other methods. These lower stock levels then relate clearly to the lower associated holding costs. These results are perhaps unexpected due to the poor performance of both of these methods in terms of forecasting accuracy on the MAN data set predictions, but are inherent to the way these methods work. The learning nature of both of these methods means that the predictions made by these methods are an attempt at recreating the patterns that the methods have understood from the data supplied to them. Those predictions may not, as was seen from the forecasting accuracy measures, necessarily be as accurate as the more straightforward methods, but may contain greater understanding of the underlying dependencies within the supplied data. With the MAN data mainly consisting of lumpy and intermittent items and the variability in the average inter-demand interval being extremely high (40.99), the other methods overestimate the overall required demand more than the MLP and LightGBM methods do. Overall, no clear winner can be deducted from the trade-off curves, meaning both methods are considered as superior for this assessment.

BRAF inventory performance

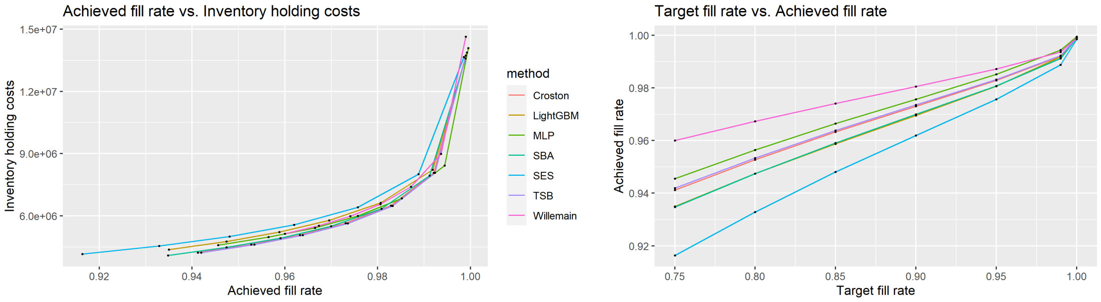
Figure 4.7 shows the result for the BRAF data set. Being classified as lumpy and intermittent, items show a high average inter-demand interval (11.14) and both items with both highly and slightly variable demand sizes. Visually, the curves resemble those of the SIM1 and SIM3 data sets. This is unusual, as the classifications of the SIM1 and SIM3 data sets were mainly erratic and smooth respectively. This means they showed a lower average inter-demand interval for all of the items. The similarity between the BRAF results and the SIM1 and SIM3 results suggests that demand variability does therefore not greatly influence the inventory control performance of the methods, providing further support for the same line of thought presented in the analysis of the SIM3 inventory control assessment, as this was the main difference between the BRAF, SIM1 and SIM3 data set characteristics. The overall superior method for this data set is again the Willemain method and all of the other methods perform very similarly. When the results are related to the forecasting accuracy results, no clear explanation for the superiority of the



(a)

(b)

Fig. 4.7 Tradeoff curves for the inventory control measures on BRAF predictions by every method.



(a)

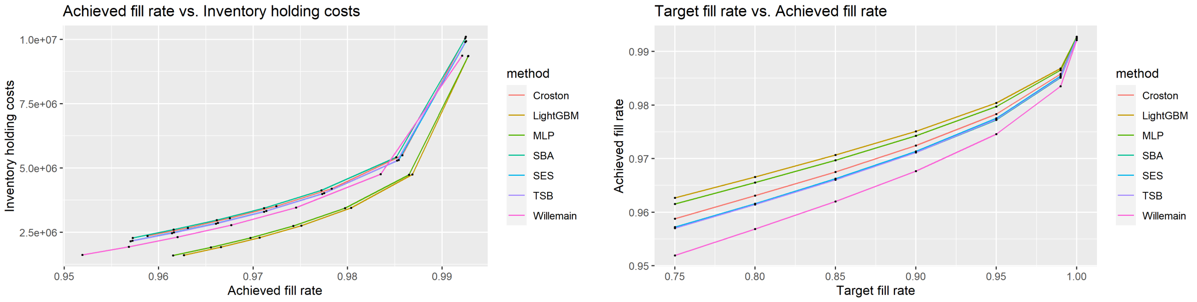
(b)

Fig. 4.8 Tradeoff curves for the inventory control measures on AUTO predictions by every method.

Willemain method seems apparent, even though based on the MSE the method performed third best.

AUTO inventory performance

With the AUTO data set being the most diverse in terms of classification, it's trade-off curves also show the most diversity in their results. The AUTO data set was classified as mostly smooth and intermittent, but also contained a portion of erratic and lumpy items. Figure 4.8a shows firstly that SES performs worst overall based on the holding costs and the achieved fill rate. This may be again accredited to the presence of erratic and lumpy items with high demand variability. The best performing methods based on the holding costs associated with the achieved fill rate are TSB and Willemain, but Willemain does achieve a significantly higher fill rate for the same target fill rate. The MLP method also achieves the second best fill rates, but seems to score average based on the associated holding costs, only performing best on the right-most portion of Figure 4.8a. This means that at a very high fill rate, the method achieves the lowest holding costs. Overall Willemain seems to be superior again, but the result based on holding costs is less clear, which may be due to the great variability in item classifications in this data set.



(a) (b)

Fig. 4.9 Tradeoff curves for the inventory control measures on OIL predictions by every method.

OIL inventory performance

Then finally the OIL data set, which was classified as having around 30% lumpy and around 70% intermittent items, meaning a high average inter-demand interval for all of the items (25.75). Figure 4.9 shows similarities to the results from the MAN data set trade-off curves, with the MLP and LightGBM methods being superior. In this case, the LightGBM method also outperforms the MLP method in every instance, achieving a higher fill rate at each target fill rate and achieving lower inventory holding costs at each achieved fill rate. The achieved fill rate for the Willemain method is lowest for every target fill rate in this application, but this results in the third best performance based on the inventory holding costs. This means that although the MLP and LightGBM methods are still overall superior to the Willemain method for this data set, the Willemain method does achieve lower inventory holding costs than all of the other methods at the same achieved fill rate. The superiority of the learning methods again shows that for data with extremely high intermittency, the understanding of the underlying dependencies seems to be superior to the more straightforward estimation of the other methods.

Chapter 5

Findings, literature comparison, conclusion & discussion

In order to structure the results from the forecasting accuracy and inventory control assessments, several findings will be presented and subsequently elaborated upon. A comparison with the results from existing literature then concludes the findings. After all of the findings have been presented, the findings will be related to the initial research questions as proposed in the introduction of this paper. After attempting to answer the research questions, a conclusion with remarks on this paper and suggestions for further research are presented.

5.1 Findings

Finding 1: The pre-processing of the data may influence the forecasting accuracy performance of spare parts demand forecasting methods.

Although the focus of this paper is on comparing spare parts demand forecasting methods, preparing the data sets for use in forecasting and the inventory control assessment requires data wrangling. In the case of this paper, this proved to matter, where the first approach attempted to recreate a practical application where the training data is updated using the forecasts made by the methods. This is simple to implement in practice, but the results showed that the results for some methods may improve by instead updating the training data with the real demand values. In practice, this would imply that Lead-time demand forecasts beyond the lead time are not possible with this type of updating. The way the data is prepared may also influence the calculation time of forecasting methods. Overall, Finding 1 suggests that the way data is prepared may be of importance.

Finding 2: The ease of implementation for the assessed methods differs in terms of total run time and required knowledge.

The total run time and average run time per item were presented in 4.6 in Chapter 4. TSB showed the lowest run time, which can be accredited to the different cost function used compared to the similar Croston, SES and SBA methods. The Willemain method had the longest run time, which is due to the repeating nature of a bootstrapping method. In this paper, the bootstrapping was done 1000 times for every application, but the run time would be even longer if the amount of bootstraps would be increased. The rest of the methods are all fairly reasonable in the total run time, with all of them being done in six hours or less. In terms of ease of implementation, the differences are larger. The Croston, SES, SBA, TSB and MLP methods are highly documented and straightforward in their application, although the MLP method offers more configurability through it's hyper parameters. This means that the method is easy to implement, yet hard to perfect. The Willemain method, although well documented by it's creator, requires knowledge of certain mathematical principles such as transition matrices and Markov chains that may not be familiar with intended users of this method. Combined with the long total run time, this decreased the ease of implementation in this paper. Lastly, the LightGBM method, which may suffer slightly from it's newness. Due to the high complexity and the extent to which the method can be adapted and configured, combined with less implementation examples than the other methods, was the most difficult to implement out of all the included methods. Overall, Finding 2 suggests that the decision to select a method should also be influenced by it's run time and ease of implementation.

Finding 3: Based on the Percentage Better comparison, SBA performed the best overall with both the initial and revised approach and LightGBM the worst.

With the highest average Percentage Better scores in both Table 4.7 and Table 4.9, SBA performed the best based on forecasting accuracy for both the industrial and simulated data sets. With the method also being easy to implement and average in total run time compared to the other methods, SBA could be a great overall method for practical applications when forecasting accuracy is the required metric. The results did not carry over to the inventory control assessment, which suggests that the method may overestimate the required demand, resulting in average or below-average inventory control performance. The LightGBM method performed worst overall in both approaches, suggesting that the method is inferior when an accurate forecast is desired, with the current setup of the hyper parameters. If forecasting accuracy is the only important metric, the method could likely be adapted to suit those needs more so than in the application in this paper.

Finding 4: Based on the inventory control assessment of the SIM1, SIM3 and BRAF data sets, demand size variability seems to influence the inventory control performance of all methods to a lesser extent than the average inter-demand interval.

As three data sets with differing demand variability metrics showed similar results, their difference did not seem to influence the results. This may suggest that for the methods used in this comparison, the average inter-demand interval is a more important metric to assess whether method performance based on inventory control will be influenced for a particular data set.

Finding 5: The inventory control performance of the SES method is negatively influenced by items with highly variable demand sizes.

The assessment on inventory control for the SIM2, SIM4 and AUTO data sets suggests that the performance of the SES method is negatively influenced by the presence of lumpy and erratic items, thus being sensitive to highly variable demand sizes. As mentioned earlier, this may be accredited to the averaging nature of the method, which is influenced by great differences in demand size more than the other methods.

Finding 6: Based on inventory control performance, Willemain performs the best, except for data with extremely high intermittency.

With the Willemain method performing the best on 6 out of 8 data sets, it's inventory control performance is considered the best overall. The superiority of the method on inventory control may be accredited to it's inner workings. The method first establishes whether the next demand occurrence is likely to be positive or zero, and in doing so it in a way learns from the data, similarly to the MLP and LightGBM methods. The expectation of the following period(s) is unique to this method only, and using this method's approach in combination with a more accurate forecasting method for establishing the expected demand sizes may be an interesting new research direction. This would also decrease the method's high total run time as the jittering would no longer be required, which takes up most of the run time.

Finding 7: For data with extremely high intermittency, the MLP and LightGBM methods perform the best based on the applied inventory control measures.

For the MAN and OIL data sets, the MLP and LightGBM methods proved to be superior in terms of inventory control. As explained earlier, the way these methods learn from underlying dependencies makes their predictions less accurate when put alongside the real values compared to the other methods, but this may make them more similar to the real underlying patterns of the data, resulting in a better inventory control performance in the case of this paper.

5.2 Comparison with the results from existing literature.

Comparison with Pinçe et al. (2021).

As the literature review in this paper was built upon the framework by Pinçe et al. (2021), a comparison with the results from their work and this paper seems fitting. In their work they did a quantitative literature analysis, meaning they did not apply any of the methods they reviewed themselves, but reviewed the results of other papers and quantified the results of those papers. In their comparison of 53 papers they first started with a comparison between Croston and SBA, where SBA performed better than Croston in 85.7% of the 20 comparisons based on forecasting accuracy measures. In this paper, out of 24 comparisons, Croston only performed better once, resulting in an even greater 95,83%. Pinçe et al. (2021) mentioned in their comparison that their only occurrence of Croston outperforming SBA was when applied to a fashion sector data set. This could be the reason why the percentage is even greater in this paper's comparison, as only spare parts data was used.

When compared on inventory control measures, Pinçe et al. (2021) indicate that Croston seems to outperform SBA more often, with 45% and 21% respectively. However, they mention that the results are inconclusive in 34.1% of the comparisons. When the trade-off curves and inventory performance results in the Appendix are examined for Croston and SBA, Croston does achieve lower fill rates for every target fill rate, but the differences are so minimal that one could also interpret these results as inconclusive. However, the fact that SBA did not outperform Croston once does suggest Croston is slightly more favoured when it comes to inventory control. Pinçe et al. (2021) accredit this to the slight positive bias in the Croston method, leading to higher inventory levels, which explains the slightly higher achieved fill rates.

The next comparison done by Pinçe et al. (2021) was between the performance of Croston and SBA and the "traditional methods", under which they understand methods such as SES, Naïve and zero-forecasting. As only SES was incorporated in this paper, no real comparison of their results to those of this paper can be made here, as their results were aggregated for all of the traditional methods. The same aggregation of comparative results is done for the group of "newer methods", under which TSB, a modified version of Croston and others are understood, and for the non-parametric methods, under which the Willemain method used in this paper falls. Therefore, no further comparison with the quantitative results from Pinçe et al. (2021) is possible.

Comparison with Spiliotis et al. (2020).

Another paper with a comparable setup was the one by Spiliotis et al. (2020), specifically because the MLP approach incorporated in their paper was very similar to the one in this paper. They also incorporated SES, Croston, SBA and TSB. Another interesting method is the Gradient Boosting Trees (GBT) method they incorporated, as the LightGBM method used in this paper functions on the same principles. One of their forecasting accuracy measures was the RMSSE, which is also used in this paper. All of these similarities make the comparison of their results with those of this paper more approachable. The two major differences are that they applied the methods to only one data set and that they did not consider inventory control performance. The data set they used contains data about sales of various consumption goods, which consists of mostly smooth and erratic items according to their classification. This means the AUTO, SIM1 and SIM3 data sets are most similar to their data out of all the data sets used in this paper.

The findings presented by Spiliotis et al. (2020) show that the four best performing methods (on forecasting accuracy) are machine learning methods. The GBT method used in their comparison was one of those four. This suggests that the line of thought that is presented throughout this paper that these advanced methods require advanced knowledge on how to configure their hyper parameters but that they show great potential when understood, may indeed be a realistic assessment. However, similarly to this paper, they found that the MLP approach was outperformed by the statistical methods on forecasting accuracy, measured by the RMSSE. This is remarkable, as they applied the MLP in a series-by-series fashion, which could be expected to show improved results compared to the aggregated input approach used in this paper. Also, SBA was found by Spiliotis et al. (2020) to be the most accurate amongst the statistical methods when forecasting accuracy is concerned, similarly to the results of this paper, where SBA was consistently one of the top performing methods on forecasting accuracy.

5.3 Conclusion

The implications of the findings will now be discussed in relation to the initial questions posed by this paper in the introduction, the first of which was: *Which type of method is best for which type of data?*

Based on the findings and the literature review, selecting the appropriate type of method for the type of data will differ. In this research several types of methods have been assessed, such as parametric methods, a bootstrapping method and machine learning methods. The results showed that none of these methods performed best across both forecasting accuracy and inventory control and that the parametric method SBA proved to be the best overall in terms of forecasting accuracy and the Willemain method in terms of inventory control. When extreme intermittency was present, the MLP and LightGBM methods performed best in terms of inventory control performance.

The second question of importance to this research was: *Does the performance of the method depend on the performance measure used or on the data set to which it is applied?*

In determining whether the performance depends on the measure or the type of data set, another factor of importance was discovered during the process of this paper. As the data needs to be pre-processed before the methods are applied, this may influence the results, as the revision of the approaches showed. Next to that, the performance of a method does depend also on the measure used, as there was no clearly superior method for some of the forecasting accuracy performance measures and inventory control measures. Similarly, there was no clearly superior method based on the data sets they were applied to, which suggests that the data does indeed influence the performance of each method. The severity of this influence is however not part of the scope of this research.

5.4 Discussion

As the process of this research showed through the revision of the approaches that the field of forecasting (spare parts demand) is a field of trial and error, there are always more options and approaches to be discovered. As some decisions made in this research have pushed the results and findings in a certain direction, so too would the implementation of more methods, different data or a different scope. In order to establish what this research suggests as further research possibilities and what was left out of the scope of this research, some possibilities will now be discussed.

As the data sets, specifically the industrial data sets, contained highly variable demand sizes and pricing, one of the initial lines of thought is to further investigate whether a weight based on price would be beneficial. A measure similar to the WRMSSE, used in the M5 competition by Makridakis et al. (2020b), would elaborate on the performance of a method when pricing differs greatly, as was the case for the industrial data sets. However, since the inventory control measures applied in this paper already incorporate inventory holding costs, this was not done for this paper.

Also, some choices regarding the pre-processing of the data could have been made differently and might be made differently for future research in order to investigate the implications. The main change that would be logical is to incorporate the actual lead times instead of setting the lead time to 1. Some more of these may include, but are not limited to; removing any demand occurrences which can be attributed to planned maintenance or which can be marked as extreme outliers, splitting the data based on classification, involving installed base forecasting, looking at demand patterns such as seasonality or heavily increasing or decreasing trends or evaluating the methods' performance on items that become obsolete. Similarly, the simulated data sets could be simulated with more extreme inter-demand intervals and a larger variation in demand size, as the industrial data sets showed that this may influence method performance.

Another possible direction for further research would be to elaborate on the hyper parameters of the MLP and LightGBM methods used in this paper and to establish whether the setting of these parameters could be automated or optimised in a way that would make the methods simple in their use. The potential that both methods showed in this paper could be further investigated if the method would be more easily approachable by those who require their methods to be easily implemented.

Reference list

Ala-Risku, T. et al. (2009). Installed base information: Ensuring customer value and profitability after the sale.

Aronis, K.-P., Magou, I., Dekker, R., and Tagaras, G. (2004). Inventory control of spare parts using a bayesian approach: A case study. *European journal of operational research*, 154(3):730–739.

Auramo, J. and Ala-Risku, T. (2005). Challenges for going downstream. *International Journal of Logistics: Research and Applications*, 8(4):333–345.

Babai, M. Z., Ali, M. M., and Nikolopoulos, K. (2012). Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. *Omega*, 40(6):713–721.

Babai, M. Z., Dallery, Y., Boubaker, S., and Kalai, R. (2019). A new method to forecast intermittent demand in the presence of inventory obsolescence. *International Journal of Production Economics*, 209:30–41.

Babai, M. Z., Syntetos, A., and Teunter, R. (2014). Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence. *International Journal of Production Economics*, 157:212–219.

Bartezzaghi, E., Verganti, R., and Zotteri, G. (1999). A simulation framework for forecasting uncertain lumpy demand. *International Journal of Production Economics*, 59(1-3):499–510.

Bergmeir, C. N., Benítez Sánchez, J. M., et al. (2012). Neural networks in r using the stuttgart neural network simulator: Rsnns. American Statistical Association.

Bookbinder, J. H. and Lordahl, A. E. (1989). Estimation of inventory re-order levels using the bootstrap statistical procedure. *IIE transactions*, 21(4):302–312.

Borchers, H. W. and Karandikar, H. (2006). A data warehouse approach for estimating and characterizing the installed base of industrial products. In *2006 International Conference on Service Systems and Service Management*, volume 1, pages 53–59. IEEE.

Boutselis, P. and McNaught, K. (2019). Using bayesian networks to forecast spares demand from equipment failures in a changing service logistics context. *International Journal of Production Economics*, 209:325–333.

Boylan, J. E. and Babai, M. Z. (2016). On the performance of overlapping and non-overlapping temporal demand aggregation approaches. *International Journal of Production Economics*, 181:136–144.

Boylan, J. E. and Syntetos, A. A. (2010). Spare parts management: a review of forecasting research and extensions. *IMA journal of management mathematics*, 21(3):227–237.

Boylan, J. E., Syntetos, A. A., and Karakostas, G. C. (2008). Classification for forecasting and stock control: a case study. *Journal of the operational research society*, 59(4):473–481.

Brockhoff, K. K. and Rao, V. R. (1993). Toward a demand forecasting model for preannounced new technological products. *Journal of Engineering and Technology Management*, 10(3):211–228.

Callioni, G., de Montgros, X., Slagmulder, R., Van Wassenhove, L. N., and Wright, L. (2005). Inventory-driven costs. *harvard business review*, 83(3):135–141.

Cohen, M., Kamesam, P. V., Kleindorfer, P., Lee, H., and Tekerian, A. (1990). Optimizer: Ibm's multi-echelon inventory system for managing service logistics. *Interfaces*, 20(1):65–82.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3):289–303.

de Haan, D. (2021). GitHub repository for benchmarking spare parts demand forecasting for intermittent demand. Available at https://github.com/danieldehaan96/spdf.

Dekker, R. (2000). Voorraadbeheersing van reservedelen: een overzicht. *Praktijkboek Magazijnen Distributiecentra.*, pages 1–27.

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):536–545.

do Rego, J. R. and De Mesquita, M. A. (2015). Demand forecasting and inventory control: A simulation study on automotive spare parts. *International Journal of Production Economics*, 161:1–16.

Durlinger, P. and Paul, I. (2012). Inventory and holding costs. *Durlinger Consultant*, page 1.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *URL http://www. jstor. org/stable/2958830*.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Fildes, R. and Goodwin, P. (2007). Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6):570–576.

Fortuin, L. and Martin, H. (1999). Control of service parts. *International Journal of Operations & Production Management*.

Franses, P. H. and Legerstee, R. (2010). Do experts' adjustments on model-based sku-level forecasts improve forecast quality? *Journal of Forecasting*, 29(3):331–340.

Gardner Jr, E. S. (2006). Exponential smoothing: The state of the art—part ii. *International journal of forecasting*, 22(4):637–666.

Ghobbar, A. A. and Friend, C. H. (2002). Sources of intermittent demand for aircraft spare parts within airline operations. *Journal of Air Transport Management*, 8(4):221–231.

Ghobbar, A. A. and Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. *Computers & Operations Research*, 30(14):2097–2114.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, 30(2):127–135.

Guo, F., Diao, J., Zhao, Q., Wang, D., and Sun, Q. (2017). A double-level combination approach for demand forecasting of repairable airplane spare parts based on turnover data. *Computers & Industrial Engineering*, 110:92–108.

Gutierrez, R. S., Solis, A. O., and Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. *International journal of production economics*, 111(2):409–420.

Hasni, M., Aguir, M., Babai, M., and Jemai, Z. (2019). Spare parts demand forecasting: a review on bootstrapping methods. *International Journal of Production Research*, 57(15-16):4791–4804.

Hua, Z., Zhang, B., Yang, J., and Tan, D. (2007). A new approach of forecasting intermittent demand for spare parts inventories in the process industries. *Journal of the Operational Research Society*, 58(1):52–61.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Jalil, M. N., Zuidwijk, R. A., Fleischmann, M., and Van Nunen, J. A. (2011). Spare parts logistics and installed base information. *Journal of the operational Research Society*, 62(3):442–457.

Jiang, A., Tam, K. L., Guo, X., and Zhang, Y. (2020). A new approach to forecasting intermittent demand based on the mixed zero-truncated poisson model. *Journal of Forecasting*, 39(1):69–83.

Johnston, F. and Boylan, J. E. (1996). Forecasting for items with intermittent demand. *Journal of the operational research society*, 47(1):113–121.

Kaggle (2020). M5 Forecasting - Accuracy. https://www.kaggle.com/c/m5-forecasting-accuracy.

Kailex (2020). M5 ForecasteR v2. Kaggle. https://www.kaggle.com/kailex/m5-forecaster-v2.

Kim, T. Y., Dekker, R., and Heij, C. (2017). Spare part demand forecasting for consumer goods using installed base information. *Computers & Industrial Engineering*, 103:201–215.

Klassen, R. D. and Flores, B. E. (2001). Forecasting practices of canadian firms: Survey results and comparisons. *International journal of production economics*, 70(2):163–174.

Kocer, U. U. (2013). Forecasting intermittent demand by markov chain model. *International Journal of Innovative Computing, Information and Control*, 9(8):3307–3318.

Kostenko, A. V. and Hyndman, R. J. (2006). A note on the categorization of demand patterns. *Journal of the Operational Research Society*, 57(10):1256–1257.

Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. *International Journal of Production Economics*, 143(1):198–206.

Kourentzes, N. (2014). On intermittent demand model optimisation and selection. *International Journal of Production Economics*, 156:180–190.

Lengu, D., Syntetos, A. A., and Babai, M. Z. (2014). Spare parts management: Linking distributional assumptions to demand classification. *European Journal of Operational Research*, 235(3):624–635.

Li, C. and Lim, A. (2018). A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. *European Journal of Operational Research*, 269(3):860–869.

Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., and Gucci, S. (2017). Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics*, 183:116–128.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting*, 1(2):111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., and Simmons, L. F. (1993). The m2-competition: A real-time judgmentally based forecasting study. *International Journal of forecasting*, 9(1):5–22.

Makridakis, S. and Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476.

Makridakis, S., Hibon, M., Lusk, E., and Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the m-competition. *International Journal of Forecasting*, 3(3-4):489–508.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020a). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020b). The m5 accuracy competition: Results, findings and conclusions. *Int J Forecast*.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (2008). *Forecasting methods and applications*. John wiley & sons.

Mathews, B. P. and Diamantopoulos, A. (1986). Managerial intervention in forecasting. an empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, 3(1):3–10.

Mathews, B. P. and Diamantopoulos, A. (1992). Judgemental revision of sales forecasts: The relative performance of judgementally revised versus non-revised forecasts. *Journal of Forecasting*, 11(6):569–576.

McCarthy, T. M., Davis, D. F., Golicic, S. L., and Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting*, 25(5):303–324.

Microsoft (2021). Welcome to LightGBM's documentation! — LightGBM 3.2.1.99 documentation. https://lightgbm.readthedocs.io/en/latest/index.html.

Mohammadipour, M. and Boylan, J. E. (2012). Forecast horizon aggregation in integer autoregressive moving average (inarma) models. *Omega*, 40(6):703–712.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moon, S., Hicks, C., and Simpson, A. (2012). The development of a hierarchical forecasting method for predicting spare parts demand in the south korean navy—a case study. *International Journal of Production Economics*, 140(2):794–802.

Mukhopadhyay, S., Solis, A. O., and Gutierrez, R. S. (2012). The accuracy of non-traditional versus traditional methods of forecasting lumpy demand. *Journal of Forecasting*, 31(8):721–735.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., and Assimakopoulos, V. (2011). An aggregate–disaggregate intermittent demand approach (adida) to forecasting: an empirical proposition and analysis. *Journal of the Operational Research Society*, 62(3):544–554.

O'Connor, M., Remus, W., and Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, 9(2):163–172.

Pennings, C. L., Van Dalen, J., and van der Laan, E. A. (2017). Exploiting elapsed time for managing intermittent demand for spare parts. *European Journal of Operational Research*, 258(3):958–969.

Petropoulos, F., Fildes, R., and Goodwin, P. (2016a). Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, 249(3):842–852.

Petropoulos, F. and Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, 66(6):914–924.

Petropoulos, F., Kourentzes, N., and Nikolopoulos, K. (2016b). Another look at estimators for intermittent demand. *International Journal of Production Economics*, 181:154–161.

Petropoulos, F. and Makridakis, S. (2020). The m4 competition: Bigger. stronger. better. *International Journal of Forecasting*, 36(1):3–6.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K. (2014). 'horses for courses' in demand forecasting. *European Journal of Operational Research*, 237(1):152–163.

Petrović, D. and Petrović, R. (1992). Sparta ii: Further development in an expert system for advising on stocks of spare parts. *International journal of production economics*, 24(3):291–300.

Pinçe, Ç., Turrini, L., and Meissner, J. (2021). Intermittent demand forecasting for spare parts: A critical review. *Omega*, page 102513.

Porras, E. and Dekker, R. (2008). An inventory control system for spare parts at a refinery: An empirical comparison of different re-order point methods. *European Journal of Operational Research*, 184(1):101–132.

Prestwich, S. D., Tarim, S. A., Rossi, R., and Hnich, B. (2014). Forecasting intermittent demand by hyperbolic-exponential smoothing. *International Journal of Forecasting*, 30(4):928–933.

Romeijnders, W., Teunter, R., and Van Jaarsveld, W. (2012). A two-step method for forecasting spare parts demand using information on component repairs. *European Journal of Operational Research*, 220(2):386–393.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., and Ducq, Y. (2013). Demand forecasting by temporal aggregation. *Naval Research Logistics (NRL)*, 60(6):479–498.

Sanders, N. R. and Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega*, 31(6):511–522.

Sanders, N. R. and Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, 5(1):39–52.

Sanders, N. R. and Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In *Principles of forecasting*, pages 405–416. Springer.

Sani, B. and Kingsman, B. G. (1997). Selecting the best periodic inventory control and demand forecasting methods for low demand items. *Journal of the operational research society*, 48(7):700–713.

Seifert, M., Siemsen, E., Hadida, A. L., and Eisingerich, A. B. (2015). Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36:33–45.

Smith, M. and Babai, M. Z. (2011). A review of bootstrapping for spare parts forecasting. *Service parts management*, pages 125–141.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85.

Snyder, R. (2002). Forecasting sales of slow and fast moving inventories. *European Journal of Operational Research*, 140(3):684–699.

Snyder, R. D., Ord, J. K., and Beaumont, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, 28(2):485–496.

Spiliotis, E., Makridakis, S., Semenoglou, A.-A., and Assimakopoulos, V. (2020). Comparison of statistical and machine learning methods for daily sku demand forecasting. *Operational Research*, pages 1–25.

Syntetos, A. A., Babai, M. Z., and Gardner Jr, E. S. (2015). Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. *Journal of Business Research*, 68(8):1746–1752.

Syntetos, A. A., Babai, M. Z., Lengu, D., and Altay, N. (2011). Distributional assumptions for parametric forecasting of intermittent demand. In *Service Parts Management*, pages 31–52. Springer.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., and Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1):1–26.

Syntetos, A. A. and Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of forecasting*, 21(2):303–314.

Syntetos, A. A. and Boylan, J. E. (2006). On the stock control performance of intermittent demand estimators. *International Journal of Production Economics*, 103(1):36–47.

Syntetos, A. A., Boylan, J. E., and Croston, J. (2005). On the categorization of demand patterns. *Journal of the operational research society*, 56(5):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):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):72–81.

Teunter, R. H. and Duncan, L. (2009). Forecasting intermittent demand: a comparative study. *Journal of the Operational Research Society*, 60(3):321–329.

Teunter, R. H., Syntetos, A. A., and Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606–615.

Turner, D. S. (1990). The role of judgement in macroeconomic forecasting. *Journal of Forecasting*, 9(4):315–345.

Turrini, L. and Meissner, J. (2019). Spare parts inventory management: New evidence from distribution fitting. *European Journal of Operational Research*, 273(1):118–130.

Van Jaarsveld, W. and Dekker, R. (2011). Estimating obsolescence risk from demand data to enhance inventory control—a case study. *International Journal of Production Economics*, 133(1):423–431.

Van Wingerden, E., Basten, R. J. I., Dekker, R., and Rustenburg, W. (2014). More grip on inventory control through improved forecasting: A comparative study at three companies. *International journal of production economics*, 157:220–237.

Viswanathan, S. and Zhou, C. (2008). A new bootstrapping based method for forecasting and safety stock determination for intermittent demand items. In *Nanyang Business School, Nanyang Technological University Singapore Working paper*.

Wang, M.-C. and Rao, S. S. (1992). Estimating reorder points and other management science applications by bootstrap procedure. *European journal of operational research*, 56(3):332–342.

Wang, W. and Syntetos, A. A. (2011). Spare parts demand: Linking forecasting to equipment maintenance. *Transportation Research Part E: Logistics and Transportation Review*, 47(6):1194–1209.

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):375–387.

Willemain, T. R., Smart, C. N., Shockor, J. H., and DeSautels, P. A. (1994). Forecasting intermittent demand in manufacturing: a comparative evaluation of croston's method. *International journal of forecasting*, 10(4):529–538.

Williams, T. (1984). Stock control with sporadic and slow-moving demand. *Journal of the Operational Research Society*, 35(10):939–948.

Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62.

Zhou, C. and Viswanathan, S. (2011). Comparison of a new bootstrapping method with parametric approaches for safety stock determination in service parts inventory systems. *International Journal of Production Economics*, 133(1):481–485.

Zhu, S., Dekker, R., Van Jaarsveld, W., Renjie, R. W., and Koning, A. J. (2017). An improved method for forecasting spare parts demand using extreme value theory. *European Journal of Operational Research*, 261(1):169–181.

Zhu, S., van Jaarsveld, W., and Dekker, R. (2020). Spare parts inventory control based on maintenance planning. *Reliability Engineering & System Safety*, 193:106600.

Appendix A

Table A.1 Inventory performance measures for the SIM1 data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9423585	44694652	0.7500000	Croston
0.9536177	48733699	0.8000000	Croston
0.9638693	53441702	0.8500000	Croston
0.9734651	59365445	0.9000000	Croston
0.9829830	68145324	0.9500000	Croston
0.9923018	84614916	0.9900000	Croston
0.9994298	143269791	0.9999999	Croston
0.9423605	44695137	0.7500000	SES
0.9536194	48734184	0.8000000	SES
0.9638708	53442187	0.8500000	SES
0.9734658	59365930	0.9000000	SES
0.9829838	68145809	0.9500000	SES
0.9923022	84615401	0.9900000	SES
0.9994298	143270276	0.9999999	SES
0.9391145	43888680	0.7500000	SBA
0.9509275	47927727	0.8000000	SBA
0.9616884	52635730	0.8500000	SBA
0.9717933	58559472	0.9000000	SBA
0.9818950	67339351	0.9500000	SBA
0.9917740	83808943	0.9900000	SBA
0.9993818	142463818	0.9999999	SBA
0.9424799	44702360	0.7500000	TSB
0.9537281	48741407	0.8000000	TSB
0.9639666	53449410	0.8500000	TSB
0.9735429	59373152	0.9000000	TSB

Continued on next page

Table A.1 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9830374	68153031	0.9500000	TSB
0.9923294	84622624	0.9900000	TSB
0.9994307	143277498	0.9999999	TSB
0.9428531	46346369	0.7500000	MLP
0.9537746	50385416	0.8000000	MLP
0.9638207	55093419	0.8500000	MLP
0.9732898	61017161	0.9000000	MLP
0.9828143	69797040	0.9500000	MLP
0.9923465	86266632	0.9900000	MLP
0.9996179	144921507	0.9999999	MLP
0.9411354	45499589	0.7500000	LightGBM
0.9524274	49538636	0.8000000	LightGBM
0.9627800	54246639	0.8500000	LightGBM
0.9725491	60170382	0.9000000	LightGBM
0.9823092	68950261	0.9500000	LightGBM
0.9920430	85419853	0.9900000	LightGBM
0.9995110	144074728	0.9999999	LightGBM
0.9481764	47663854	0.7500000	Willemain
0.9579242	51702901	0.8000000	Willemain
0.9669104	56410904	0.8500000	Willemain
0.9753960	62334647	0.9000000	Willemain
0.9839539	71114525	0.9500000	Willemain
0.9925306	87584118	0.9900000	Willemain
0.9994058	146238992	0.9999999	Willemain

Table A.2 Inventory performance measures for the SIM2 data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9335466	15738399	0.7500000	Croston
0.9462346	17526649	0.8000000	Croston
0.9577866	19611074	0.8500000	Croston
0.9686105	22233755	0.9000000	Croston
0.9794861	26120965	0.9500000	Croston
0.9904060	33412723	0.9900000	Croston

Continued on next page

Table A.2 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9991610	59381621	0.9999999	Croston
0.8985701	15579080	0.7500000	SES
0.9181696	17367330	0.8000000	SES
0.9359649	19451755	0.8500000	SES
0.9525734	22074436	0.9000000	SES
0.9692254	25961646	0.9500000	SES
0.9857250	33253404	0.9900000	SES
0.9987953	59222302	0.9999999	SES
0.9294388	15287381	0.7500000	SBA
0.9429361	17075631	0.8000000	SBA
0.9551942	19160055	0.8500000	SBA
0.9667105	21782737	0.9000000	SBA
0.9782731	25669946	0.9500000	SBA
0.9898503	32961705	0.9900000	SBA
0.9991142	58930603	0.9999999	SBA
0.9339912	15717537	0.7500000	TSB
0.9466470	17505788	0.8000000	TSB
0.9581647	19590212	0.8500000	TSB
0.9689514	22212894	0.9000000	TSB
0.9797794	26100103	0.9500000	TSB
0.9905895	33391861	0.9900000	TSB
0.9991832	59360760	0.9999999	TSB
0.9314509	15943684	0.7500000	MLP
0.9443465	17731934	0.8000000	MLP
0.9561724	19816358	0.8500000	MLP
0.9673409	22439040	0.9000000	MLP
0.9786498	26326250	0.9500000	MLP
0.9901615	33618008	0.9900000	MLP
0.9993525	59586906	0.9999999	MLP
0.9312231	15885510	0.7500000	LightGBM
0.9442420	17673760	0.8000000	LightGBM
0.9560656	19758184	0.8500000	LightGBM
0.9672674	22380866	0.9000000	LightGBM
0.9785850	26268075	0.9500000	LightGBM
0.9900684	33559834	0.9900000	LightGBM

Continued on next page

Table A.2 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9992793	59528732	0.9999999	LightGBM
0.9361381	16454177	0.7500000	Willemain
0.9479614	18242427	0.8000000	Willemain
0.9588437	20326851	0.8500000	Willemain
0.9691346	22949533	0.9000000	Willemain
0.9795904	26836743	0.9500000	Willemain
0.9902523	34128501	0.9900000	Willemain
0.9991153	60097399	0.9999999	Willemain

Table A.3 Inventory performance measures for the SIM3 data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9529304	34546140	0.7500000	Croston
0.9635260	36992170	0.8000000	Croston
0.9729595	39843317	0.8500000	Croston
0.9815065	43430710	0.9000000	Croston
0.9895083	48747768	0.9500000	Croston
0.9964036	58721684	0.9900000	Croston
0.9999114	94242828	0.9999999	Croston
0.9489077	34251478	0.7500000	SES
0.9604439	36697508	0.8000000	SES
0.9706760	39548655	0.8500000	SES
0.9799463	43136048	0.9000000	SES
0.9886540	48453106	0.9500000	SES
0.9961223	58427022	0.9900000	SES
0.9999036	93948167	0.9999999	SES
0.9491452	33901847	0.7500000	SBA
0.9604985	36347877	0.8000000	SBA
0.9706288	39199024	0.8500000	SBA
0.9798348	42786417	0.9000000	SBA
0.9885058	48103475	0.9500000	SBA
0.9960173	58077391	0.9900000	SBA
0.9999017	93598536	0.9999999	SBA
0.9531562	34551798	0.7500000	TSB

Continued on next page

Table A.3 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9637397	36997828	0.8000000	TSB
0.9731461	39848975	0.8500000	TSB
0.9816550	43436368	0.9000000	TSB
0.9896148	48753426	0.9500000	TSB
0.9964527	58727341	0.9900000	TSB
0.9999123	94248486	0.9999999	TSB
0.9489333	34525830	0.7500000	MLP
0.9601106	36971860	0.8000000	MLP
0.9702430	39823006	0.8500000	MLP
0.9795480	43410400	0.9000000	MLP
0.9884127	48727457	0.9500000	MLP
0.9963046	58701373	0.9900000	MLP
0.9999668	94222518	0.9999999	MLP
0.9500425	34541002	0.7500000	LightGBM
0.9610274	36987032	0.8000000	LightGBM
0.9709546	39838179	0.8500000	LightGBM
0.9800337	43425572	0.9000000	LightGBM
0.9886376	48742630	0.9500000	LightGBM
0.9961741	58716546	0.9900000	LightGBM
0.9999437	94237691	0.9999999	LightGBM
0.9604739	36753963	0.7500000	Willemain
0.9692217	39199993	0.8000000	Willemain
0.9770527	42051140	0.8500000	Willemain
0.9841753	45638534	0.9000000	Willemain
0.9908997	50955591	0.9500000	Willemain
0.9967847	60929507	0.9900000	Willemain
0.9999188	96450652	0.9999999	Willemain

Table A.4 Inventory performance measures for the SIM4 data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9405261	15089715	0.7500000	Croston
0.9556926	16485397	0.8000000	Croston
0.9688045	18112235	0.8500000	Croston

Continued on next page

Table A.4 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9801479	20159168	0.9000000	Croston
0.9900675	23193032	0.9500000	Croston
0.9974096	28884056	0.9900000	Croston
0.9999881	49152091	0.9999999	Croston
0.9109691	14993343	0.7500000	SES
0.9338213	16389025	0.8000000	SES
0.9535180	18015863	0.8500000	SES
0.9705558	20062796	0.9000000	SES
0.9853563	23096660	0.9500000	SES
0.9962302	28787683	0.9900000	SES
0.9999816	49055719	0.9999999	SES
0.9351092	14685051	0.7500000	SBA
0.9515103	16080733	0.8000000	SBA
0.9657588	17707570	0.8500000	SBA
0.9781383	19754504	0.9000000	SBA
0.9889960	22788367	0.9500000	SBA
0.9971155	28479391	0.9900000	SBA
0.9999856	48747426	0.9999999	SBA
0.9409140	15064568	0.7500000	TSB
0.9560939	16460250	0.8000000	TSB
0.9691693	18087088	0.8500000	TSB
0.9804654	20134021	0.9000000	TSB
0.9902873	23167885	0.9500000	TSB
0.9974912	28858908	0.9900000	TSB
0.9999878	49126944	0.9999999	TSB
0.9391769	15254805	0.7500000	MLP
0.9543532	16650487	0.8000000	MLP
0.9676849	18277325	0.8500000	MLP
0.9793722	20324258	0.9000000	MLP
0.9897091	23358122	0.9500000	MLP
0.9976600	29049146	0.9900000	MLP
0.9999982	49317181	0.9999999	MLP
0.9372906	15061501	0.7500000	LightGBM
0.9530006	16457183	0.8000000	LightGBM
0.9667508	18084021	0.8500000	LightGBM

Continued on next page

Table A.4 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9787789	20130954	0.9000000	LightGBM
0.9893262	23164818	0.9500000	LightGBM
0.9973403	28855842	0.9900000	LightGBM
0.9999952	49123877	0.9999999	LightGBM
0.9471384	15897659	0.7500000	Willemain
0.9604547	17293341	0.8000000	Willemain
0.9720097	18920178	0.8500000	Willemain
0.9820799	20967112	0.9000000	Willemain
0.9908928	24000975	0.9500000	Willemain
0.9975670	29691999	0.9900000	Willemain
0.9999877	49960034	0.9999999	Willemain

Table A.5 Inventory performance measures for the MAN data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9540260	178090.9	0.7500000	Croston
0.9597055	204826.8	0.8000000	Croston
0.9651039	235990.9	0.8500000	Croston
0.9705284	275202.4	0.9000000	Croston
0.9765834	333319.7	0.9500000	Croston
0.9841197	442338.2	0.9900000	Croston
0.9944322	830596.9	0.9999999	Croston
0.9541098	173629.6	0.7500000	SES
0.9597806	200365.6	0.8000000	SES
0.9651952	231529.6	0.8500000	SES
0.9706384	270741.1	0.9000000	SES
0.9766585	328858.4	0.9500000	SES
0.9841634	437876.9	0.9900000	SES
0.9944388	826135.6	0.9999999	SES
0.9534522	175222.9	0.7500000	SBA
0.9592366	201958.9	0.8000000	SBA
0.9647304	233122.9	0.8500000	SBA
0.9702344	272334.4	0.9000000	SBA
0.9763753	330451.7	0.9500000	SBA

Continued on next page

Table A.5 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9839944	439470.2	0.9900000	SBA
0.9944038	827728.9	0.9999999	SBA
0.9539155	173596.0	0.7500000	TSB
0.9595782	200332.0	0.8000000	TSB
0.9650056	231496.0	0.8500000	TSB
0.9704527	270707.5	0.9000000	TSB
0.9765093	328824.8	0.9500000	TSB
0.9840406	437843.3	0.9900000	TSB
0.9943942	826102.1	0.9999999	TSB
0.9412888	112282.1	0.7500000	MLP
0.9498840	139018.1	0.8000000	MLP
0.9577393	170182.1	0.8500000	MLP
0.9654068	209393.6	0.9000000	MLP
0.9736056	267510.9	0.9500000	MLP
0.9830865	376529.4	0.9900000	MLP
0.9946689	764788.2	0.9999999	MLP
0.9416146	112653.2	0.7500000	LightGBM
0.9501792	139389.2	0.8000000	LightGBM
0.9579536	170553.2	0.8500000	LightGBM
0.9655950	209764.7	0.9000000	LightGBM
0.9737475	267882.0	0.9500000	LightGBM
0.9831610	376900.5	0.9900000	LightGBM
0.9946815	765159.3	0.9999999	LightGBM
0.9548907	174486.9	0.7500000	Willemain
0.9603536	201222.8	0.8000000	Willemain
0.9655914	232386.9	0.8500000	Willemain
0.9708487	271598.4	0.9000000	Willemain
0.9767714	329715.7	0.9500000	Willemain
0.9841342	438734.2	0.9900000	Willemain
0.9943699	826992.9	0.9999999	Willemain

Table A.6 Inventory performance measures for the BRAF data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9540260	178090.9	0.7500000	Croston
0.9597055	204826.8	0.8000000	Croston
0.9651039	235990.9	0.8500000	Croston
0.9705284	275202.4	0.9000000	Croston
0.9765834	333319.7	0.9500000	Croston
0.9841197	442338.2	0.9900000	Croston
0.9944322	830596.9	0.9999999	Croston
0.9541098	173629.6	0.7500000	SES
0.9597806	200365.6	0.8000000	SES
0.9651952	231529.6	0.8500000	SES
0.9706384	270741.1	0.9000000	SES
0.9766585	328858.4	0.9500000	SES
0.9841634	437876.9	0.9900000	SES
0.9944388	826135.6	0.9999999	SES
0.9534522	175222.9	0.7500000	SBA
0.9592366	201958.9	0.8000000	SBA
0.9647304	233122.9	0.8500000	SBA
0.9702344	272334.4	0.9000000	SBA
0.9763753	330451.7	0.9500000	SBA
0.9839944	439470.2	0.9900000	SBA
0.9944038	827728.9	0.9999999	SBA
0.9539155	173596.0	0.7500000	TSB
0.9595782	200332.0	0.8000000	TSB
0.9650056	231496.0	0.8500000	TSB
0.9704527	270707.5	0.9000000	TSB
0.9765093	328824.8	0.9500000	TSB
0.9840406	437843.3	0.9900000	TSB
0.9943942	826102.1	0.9999999	TSB
0.9412888	112282.1	0.7500000	MLP
0.9498840	139018.1	0.8000000	MLP
0.9577393	170182.1	0.8500000	MLP
0.9654068	209393.6	0.9000000	MLP
0.9736056	267510.9	0.9500000	MLP
0.9830865	376529.4	0.9900000	MLP

Continued on next page

Table A.6 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9946689	764788.2	0.9999999	MLP
0.9416146	112653.2	0.7500000	LightGBM
0.9501792	139389.2	0.8000000	LightGBM
0.9579536	170553.2	0.8500000	LightGBM
0.9655950	209764.7	0.9000000	LightGBM
0.9737475	267882.0	0.9500000	LightGBM
0.9831610	376900.5	0.9900000	LightGBM
0.9946815	765159.3	0.9999999	LightGBM
0.9548907	174486.9	0.7500000	Willemain
0.9603536	201222.8	0.8000000	Willemain
0.9655914	232386.9	0.8500000	Willemain
0.9708487	271598.4	0.9000000	Willemain
0.9767714	329715.7	0.9500000	Willemain
0.9841342	438734.2	0.9900000	Willemain
0.9943699	826992.9	0.9999999	Willemain

Table A.7 Inventory performance measures for the AUTO data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9412320	4229905	0.7500000	Croston
0.9527760	4619348	0.8000000	Croston
0.9632684	5073292	0.8500000	Croston
0.9730921	5644457	0.9000000	Croston
0.9828719	6491010	0.9500000	Croston
0.9921214	8079001	0.9900000	Croston
0.9989867	13734482	0.9999999	Croston
0.9163144	4155608	0.7500000	SES
0.9329021	4545051	0.8000000	SES
0.9480452	4998995	0.8500000	SES
0.9619873	5570160	0.9000000	SES
0.9757491	6416713	0.9500000	SES
0.9887970	8004705	0.9900000	SES
0.9985519	13660185	0.9999999	SES
0.9347902	4087026	0.7500000	SBA

Continued on next page

Table A.7 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9474310	4476469	0.8000000	SBA
0.9590509	4930413	0.8500000	SBA
0.9699490	5501578	0.9000000	SBA
0.9808249	6348131	0.9500000	SBA
0.9912250	7936122	0.9900000	SBA
0.9989223	13591603	0.9999999	SBA
0.9419404	4224320	0.7500000	TSB
0.9533812	4613763	0.8000000	TSB
0.9638368	5067707	0.8500000	TSB
0.9735881	5638872	0.9000000	TSB
0.9832381	6485425	0.9500000	TSB
0.9923281	8073416	0.9900000	TSB
0.9990240	13728897	0.9999999	TSB
0.9455854	4582935	0.7500000	MLP
0.9564843	4972378	0.8000000	MLP
0.9664777	5426322	0.8500000	MLP
0.9757082	5997487	0.9000000	MLP
0.9852330	6844040	0.9500000	MLP
0.9944402	8432031	0.9900000	MLP
0.9995498	14087512	0.9999999	MLP
0.9349677	4370576	0.7500000	LightGBM
0.9474101	4760020	0.8000000	LightGBM
0.9587865	5213963	0.8500000	LightGBM
0.9695329	5785128	0.9000000	LightGBM
0.9806376	6631681	0.9500000	LightGBM
0.9918921	8219673	0.9900000	LightGBM
0.9992376	13875153	0.9999999	LightGBM
0.9599948	5139719	0.7500000	Willemain
0.9673563	5529163	0.8000000	Willemain
0.9741062	5983106	0.8500000	Willemain
0.9805733	6554271	0.9000000	Willemain
0.9871818	7400824	0.9500000	Willemain
0.9936691	8988816	0.9900000	Willemain
0.9989963	14644296	0.9999999	Willemain

Table A.8 Inventory performance measures for the OIL data set.

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9588103	2354818	0.7500000	Croston
0.9630705	2672656	0.8000000	Croston
0.9675253	3043136	0.8500000	Croston
0.9724465	3509283	0.9000000	Croston
0.9783135	4200185	0.9500000	Croston
0.9858017	5496201	0.9900000	Croston
0.9925227	10111838	0.9999999	Croston
0.9572204	2172833	0.7500000	SES
0.9616155	2490672	0.8000000	SES
0.9662596	2861151	0.8500000	SES
0.9713785	3327299	0.9000000	SES
0.9775097	4018201	0.9500000	SES
0.9854209	5314217	0.9900000	SES
0.9925342	9929854	0.9999999	SES
0.9572476	2281640	0.7500000	SBA
0.9615865	2599479	0.8000000	SBA
0.9661366	2969958	0.8500000	SBA
0.9711613	3436106	0.9000000	SBA
0.9772356	4127008	0.9500000	SBA
0.9851333	5423024	0.9900000	SBA
0.9924466	10038661	0.9999999	SBA
0.9570306	2147208	0.7500000	TSB
0.9614194	2465047	0.8000000	TSB
0.9660555	2835526	0.8500000	TSB
0.9711670	3301674	0.9000000	TSB
0.9773034	3992576	0.9500000	TSB
0.9852362	5288592	0.9900000	TSB
0.9924460	9904229	0.9999999	TSB
0.9615377	1593010	0.7500000	MLP
0.9655428	1910848	0.8000000	MLP
0.9697012	2281328	0.8500000	MLP
0.9742502	2747475	0.9000000	MLP
0.9797175	3438377	0.9500000	MLP
0.9864952	4734393	0.9900000	MLP

Continued on next page

Table A.8 – continued from previous page

Achieved fill rate	Inventory holding costs	Target fill rate	Method
0.9927557	9350030	0.9999999	MLP
0.9626769	1605204	0.7500000	LightGBM
0.9666015	1923043	0.8000000	LightGBM
0.9706708	2293522	0.8500000	LightGBM
0.9750957	2759670	0.9000000	LightGBM
0.9803812	3450572	0.9500000	LightGBM
0.9868413	4746588	0.9900000	LightGBM
0.9927780	9362225	0.9999999	LightGBM
0.9519112	1617924	0.7500000	Willemain
0.9568539	1935762	0.8000000	Willemain
0.9619998	2306242	0.8500000	Willemain
0.9676803	2772390	0.9000000	Willemain
0.9745522	3463291	0.9500000	Willemain
0.9834925	4759308	0.9900000	Willemain
0.9921160	9374945	0.9999999	Willemain