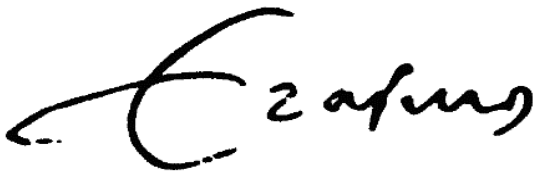


MSc THESIS

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

MSc. Data Science & Marketing Analytics

Title:

Exploring the influence of synthetic data and diverse methodologies on spare parts demand forecasting

May 1st, 2023

Kasimir Harris-Wilson

ID: 603046

Supervisor: Dr.Prof. R. Dekker

Second Assessor: Dr. Prof. ACD Donkers

ABSTRACT

This research investigates the challenges of spare parts demand forecasting in the logistics industry and explores the potential solutions using machine learning methods and synthetic data. The study evaluates the effectiveness of multiple methods and analytical pluralism and compares the benefits and drawbacks of using synthetic versus real data. The results indicate that machine learning models, particularly RNN and XGBoost models, outperform traditional statistical models like ARIMA and that using multiple methods can improve forecasting accuracy. The study also emphasizes the importance of computation power and expertise in artificial intelligence models, which are crucial for achieving more accurate results. The findings have significant implications for the spare parts industry, where the use of advanced forecasting methods can lead to improved accuracy and reduced inventory costs. Additionally, the study highlights the need for greater computing power to support more complex forecasting methods and emphasizes the importance of expertise in artificial intelligence for interpreting results accurately. Overall, this study provides valuable insights into the potential benefits of incorporating machine learning methods and synthetic data in spare parts demand forecasting, emphasizing the need for continued research in this area.

Key Terms

Spare parts demand forecasting | Machine learning methods | Synthetic data generation

Analytical pluralism | Traditional statistical models (e.g., MA) | Hybrid stacking method.

Occam's Razor | LSTM model | Data quality

Zero-value challenges | Computing power | Expertise in artificial intelligence models

Inventory costs | Supply chain | Logistics industry

Forecast improvement strategies | Intermittent data predictions | Counterfactual analysis.

Content

1. Introduction

Forecasting in logistics is crucial for supply chain companies to plan. The goal is to prevent shortages and excess orders, which can affect operating costs. By introducing new machine learning methods, the global economy can better predict unforeseen circumstances and protect supply chains. On top of forecasting, inputs of sufficient quality are needed to make coherent forecasts, but inputs may not meet quality standards let alone be available. This thesis, thus, focuses on forecasting methods and their inputs.

Concerning forecasting methods, there is a trade-off between the simplicity of forecasting methods and the need to add complexity to the models. Complexity in predictive models refers to the number of features or terms included in the predictive models. Analytical pluralism involves adding more features to the models, but it raises the question of whether it is aligned with the current needs of forecasting.[1] The opposite of analytical pluralism is the preference toward simpler models, which is parallel to the principle of parsimony, famously known as Ockham's razor (Britannica, 1998).[2] This principle, established by William of Ockham, states that explanations using fewer entities or types of entities are preferred over those using more. The principle of parsimony is widely used in areas such as forecasting and contradicts the approach of analytical pluralism. However, recent studies suggest that newer approaches have made multiple methods more predictable than stand-alone models, and this is one focus of this paper.

Analytical pluralism, on the other hand, is according to Clarke et al. (2015) the use of multiple analytical methods on a single dataset. The idea is that using multiple methods will provide a more comprehensive understanding of the phenomenon. However, the practical use of analytical pluralism can be challenging due to differences in theoretical frameworks. Despite this, analytical pluralism is used because single or simple methods are either limited to specific purposes or unable to assess multiple dimensions within a phenomenon (Frost et al., 2011; Kincheloe, 2001; 2005). To improve predictions of spare parts demand, predictive models are being used in conjunction with other models. An alternative interpretation of analytical pluralism is known as forecast improvement strategies, coined *demand characteristics* which entail applying forecasting methods based on demand characteristics (Pinçe et al. 2021; Syntetos and Boylan, 2005; Petropoulos et al., 2018).

Predictive models, furthermore, are not only meant to foretell an outcome but may also be used to enhance our inputs that have quality issues. This is because data entries are often prone to human error, which is why decision-makers may opt for artificially generated inputs, known as synthetic data, to assimilate real-life information. Synthetic inputs are said to offer several advantages over their real counterpart, such as privacy protection (Abowd and Lane, 2004). However, it is the exact benefits it adds

[1] For the sake of disambiguation, it is noted that pluralism – frequently used in the field of political science is more recently used in conjunction with machine learning (ML).

[2] "The explanation requiring the fewest assumptions is most likely to be correct." – William of Ockham

to our predictive model in terms of accuracy that needs to be better established, which is why this paper ought to investigate the differences between synthetic and real data in affecting our prediction models.

Aside from improved privacy, it is known from current literature that synthetic data offers benefits, such as improved quality compared to real data that may contain inaccuracies and biases. Synthetic data is also scalable, allowing for the generation of a large amount of information from existing data. Despite these benefits, existing literature cannot pinpoint its potential merits in improving our predictive models. What seems even less clear are the potential drawbacks of feeding our predictive models with synthetic data, which motivates this paper to further dwell on the topic.

The use of synthetic data has risen in response to growing demands for improved privacy, but there is still a lack of consensus on its merits (James et. al, 2021). This paper aims to contribute to our understanding of how synthetic and real inputs impact standalone and multiple methods. Forecasts are broad and applicable to many themes. The research ought to narrow down the focus specifically to spare parts. As the name suggests, a spare part describes parts meant to repair or replace failed units. Accurate forecasting of spare parts demand is critical for the smooth functioning of global supply chains. However, predicting the demand for spare parts is hampered by many zero values. This paper, therefore, will address three research questions:

1. *How do simple models and analytical pluralism impact spare parts demand forecasting accuracy, and what role does input data quality play?*

2. *What are the benefits and drawbacks of using synthetic versus real data in spare parts demand prediction?*

3. *How can machine learning methods enhance traditional forecasting approaches for spare parts demand, addressing data quality and zero-value challenges?*

This paper is structured as follows: the second section is a literature review, which includes an overview of research on the two competing theoretical models and inputs. The latter part of the section goes into further detail on the methods used to derive spare parts demand. The third section presents the experimental design and methodology, followed by the results in the fourth section. The paper concludes with a concluding section.

2. Literature Review

The literature review is divided into two parts. The first part examines the theories mentioned in the introduction, including spare parts, analytical pluralism, and Ockham's Razor. The literature on spare parts provides insight into the characteristics of spare part demand, which informs the methods discussed in the subsequent section of the literature review. The second part highlights methods used in past research to predict either intermittent input or spare parts demand.

Part I: Literature on Theory

The first few sections of the literature review highlight why spare parts demand is difficult to predict, going deep into the nature of intermittent data. What follows is dwell deeper into analytical pluralism applied to a spare parts demand prediction context, and dwell on its counterpart the law of parsimony, famously known as *Ockahm's Razor***.**

The following sections of the first part of the literature review also delves into the growing importance of synthetic data in modern applications, highlighting its advantages in privacy protection and cost savings. Various methods for generating synthetic data are explored, including R-language programs such as *Synthpop* and *Fabricatr*, as well as deep learning techniques like Generative Adversarial Networks (GANs) and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks.

The review acknowledges the limited research available on using RNNs for synthesizing time series data but emphasizes the potential of LSTMs in this context, given their ability to model nonlinear relationships and their success as discriminative models. By replicating a recent study that proposed a new LSTM architecture using an autoencoder, the thesis aims to evaluate the potential of LSTM as a generative model for synthesizing spare parts demand data.

2.1. The Intuition of Spare Parts Demand

Spare parts' key characteristic from a statistical point of view is that their demand can be intermittent.[3] In other words, it occurs at irregular intervals. Mathematically, we can explain intermittency as widely spaced frequencies that are concentrated in the probability distribution function (PDF) where the local PDF is substantially greater than PDF's mean, i.e. $x \gg \bar{x}$. Intermittency is further synonymous with randomness, which comes in varying shapes: mild, slow, and wild (Diamond, 2019). Interestingly, the level of randomness has different implications.

[3] Other classifications are described as lumpy, erratic or smooth (Syntetos and Boylan, 2001).

Spare part's associated irregular intervals mean that time series are subjected to many zero values, which harms forecasting methods' ability to project reliable forecasts. To demonstrate why issues, arise with irregular values, we use the example made by Waller (2015). The first method proposed to be applied to intermittent demand was single exponential smoothing (SES):

$$F_{t+1} = \alpha x_t + (1 - \alpha) F_t \quad (1)$$

Equation (1) suggests that the forecast demand in the following period is a weighted average between two quantities — where F_t is the forecast for periods t. Actual demand is denoted by x for periods t. Lastly, α describes the smoothing parameter that is adjustable between 0 and 1. The intuition follows that a higher smoothing parameter adjusts greater to changes whilst being less robust to noise. The underlying issue with this method — considering intermittency, is that it has an upward bias in the forecast in periods directly after non-zero demand. Yet, the approach originates from 1956, and fortunately, new forecasting models were developed to tackle the underlying issue. Overall, spare parts' demand characteristics pose a challenge to accurately predicting demand.

2.2. The Characteristics of Spare Parts Demand

Intermittency is a type of randomness that was originally classified as slow or wild. However, Mandelbrot (1997) expanded the statistical model by introducing the concept of "mild" randomness to better capture real-world turbulence. These three states depart from Gaussian statistics to varying degrees.[4] Slow randomness is linked to states with growing higher-order moments, mild randomness is associated with defined means and 2nd-order moment variance, and wild randomness is characterized by the absence of convergence for even the lowest-order moments (Diamond, 2019). Intermittency leads to rare but intense peaks in quantities over time, which complicates forecasts. Gaussian statistical moments require the variance to be convergent as additive components increase, but intermittent values have a multiplicative process (Diamond, 2019). To conclude, spare parts demand therefore tends to be non-parametric, but many statistical models require a parametric assumption which may invalidate their predictions.

Newer methods to adjust for randomness might differ in effectiveness as randomness comes in varying degrees. From existing research, it is inferred that spare part demand can be characterized as having lumpy demand patterns, very intermittent, having zero demand periods — or occasional zero demand periods. The many zero entries are intuitively linked with the fact spare parts are part replacements of the installed base of machines, which are either purchased correctively or preventively (Van der Auweraer et. al, 2019). From the same source, it is inferred that the installed base describes the

[4] The principles are specifically coined the Central Limit Theorem (CLT) and Law of Large Numbers (LLN).

number of products sold to lead to the need for their spare parts. Already existing literature emphasizes the need to have awareness of characteristics associated with the installed base to predominately make correct inventory decisions (Dekker et. al, 2013). Failure to account for the prior may result in excessive holding costs, and inferior spare part demand forecasts (Guvenir and Erel, 1998).

To capture differences in demand profiles, the demand is classified based on the methodology proposed by Boylan et al. (2008) and Syntetos et al. (2005). This approach captures differences in demand profiles by utilizing two parameters to derive demand regularity in quantity (CV^2) and in time (ADI). The Average Demand Interval (ADI) describes the average interval time between two demand occurrences, while the squared Coefficient of Variation (CV^2) represents the variability of demand sizes in the event of demand.

$$ADI = \frac{Total\ \ Amount\ of\ Time\ Periods}{Amount\ of\ Time\ Periods\ with\ Demand\ Greater\ than\ 0} \tag{2}$$

$$CV^2 = \left(\frac{Standard\ Deviation}{Mean}\right)^2 \tag{3}$$

Figure 1 below depicts Syntetos and Boylan's (2009) common classification scheme based on the average demand interval (ADI) and the squared coefficient of variation of demand sizes in the event of demand (CV^2). Briefly, ADI describes the average interval time between two demand occurrences and CV^2 is the standard deviation of the demand divided by the average demand for non-zero periods (Kaya et. al, 2020). Hence, varying types of intermittent demand add another layer of complexity as to how researchers ought to circumvent intermittency. This paper, therefore, acknowledges that difference demand classifications will affect predictive models differently, but it is denoted that our inputs in our analysis will be fed to models indiscriminately.

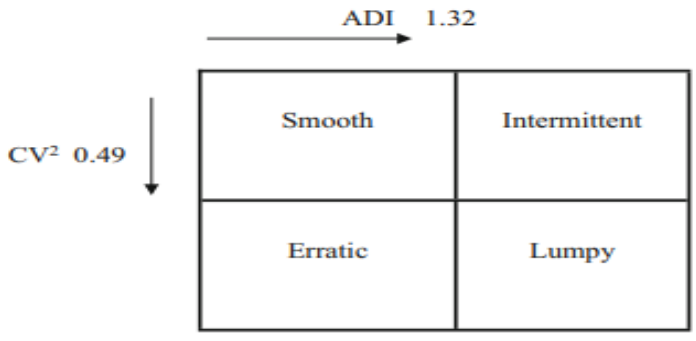


Figure 1 Boylan et al. (2008) and Syntetos et al. (2005): Categorization scheme for intermittent demand forecasting

2.3. Analytical Pluralism

In the context of analytical pluralism, Ince and Trafalis (2006) conducted a study to examine the potential benefits of incorporating an additional stage to their forecasting analysis, resulting in more complex predictive models. Meanwhile, according to Johnson et al. (2007), analytical pluralism involves combining qualitative and quantitative elements, such as viewpoints, analyses, data collection, and evaluation techniques. In the context of spare parts forecasting, Pinçe et al. (2021) distinguish between contextual and time-series forecasts, but this thesis primarily focuses on the quantitative time-series aspect. From Pinçe et al.'s paper, it can be inferred that the time-series domain includes both quantitative methods and improvement strategies that aim to improve predictions by altering or classifying inputs. In this paper, we consider forecasting improvement strategies, such as demand classification and data aggregation, as well as combined forecasts, to be a form of analytical pluralism since they involve using multiple approaches to achieve better forecasting results.

2.3.1. Combined Forecasts

Granger and Bates (1969) proposed the concept of combining forecasting methods to improve accuracy. Han et al. (2017) suggest that this can be achieved by combining information from multiple forecasting methods. They propose a combined prediction model where the true value of a prediction over duration is denoted by $V(p)$, and the outcomes of M diverse prediction approaches are denoted by $G(p;q)$. The weighted vectors for these techniques are indicated by $R(A_1, A_2, A_M)$, adhering to normalization constraints. The combined prediction model can be formulated as:

$$\hat{V}(p) = \sum_{q=1}^{M} \left(A_q * G(p;q) \right) \quad (4)$$

There are two main types of combination forecasting methods: linear and non-linear. In linear methods, two separate prediction values are fed into a model to predict the target variable (Han et al., 2017; Guitierrez et al., 2008; Mitra et al., 2022). In contrast to linear forecasting methods that assume a linear relationship between the variables being forecasted, non-linear forecasting methods account for non-linear relationships that may exist between variables. Combining two non-linear forecasting methods can help to capture more complex relationships and provide more accurate forecasts.

Seyedan and Mafakheri (2020) suggest that spreadsheet models used for demand forecasting are not scalable for large-scale data and that a firm's supply chain management cannot analyze complex problems through simple statistical methods. They propose mixed approaches, such as combined forecasting, as appropriate when demand forecasting becomes highly dimensional, involving factors such as several points of supply, different warehouses, and varied customers that a single method cannot account for

simultaneously. Mixed approaches are said to yield the best accuracy. However, a survey conducted by Ali et al. (2011) indicates that analyzers tend to opt for simpler methods to make forecasts.

In this thesis, we aim to explore the use of combined forecasting methods to enhance the accuracy of spare parts demand predictions. The motivation for such methods is to address the limitations of conventional linear methods and to address the complexity of spare parts demand forecasting, which we will discuss in the following sections.

2.3.2. Demand Classification

Du et al. (2017) noted that input selection is crucial for achieving better results with combined forecasts. This is because different input selection processes can yield different results. In the context of spare parts demand forecasting, a balanced approach is crucial, as certain demand characteristics should be predicted with the appropriate methods (Golicic et al., 2005). One such approach is demanding classification, which involves matching demand characteristics with the most appropriate predictive method to enhance forecasting and stock control (Pinçe et al., 2021).

The earliest demand classification approach can be traced back to Williams (1984), who classified SKUs based on their purchase velocity. Williams found that this demand classification scheme led to a reduction in inventory costs compared to assuming continuous demand for all products. A more contemporary demand classification scheme, which was mentioned previously, involves capturing differences in demand profiles using two parameters: CV^2 for demand regularity in quantity and ADI for demand regularity in time. Other alternative demand classification approaches also aim to achieve the best theoretical demand distribution for inventory control based on empirical observations (Pinçe et al., 2021).

Moreover, the demand distributions, it is inferred from the previously mentioned source that classification methods commonly assume that the interarrival times follow a Bernoulli or Poisson distribution. The issue with this assumption, according to Pinçe et al. is that they are unable to account for increasing failure rates due to their memoryless property and thus fail to capture actual intermittent demand patterns. Syntetots et al. (2014) accounted for the issue by investigating traditional forecasting methods for Erlang-distributed interarrival times and providing alternative demand classifications cutoff values.[5] For further context, the Erlang Distribution is frequently used in queuing theory to model waiting times in systems where events occur independently and at a constant average rate. Moreover, it is inferred from the paper that the average inter-demand interval is a useful classification criterion for the Erlang distribution assumption, but the CV^2 has a less potent explanatory power compared to earlier studies.

[5] The Erlang distribution is used to model the sum of k (shape parameter) exponentially distributed random variables with a rate parameter λ.

An alternative distribution assumption is a negative binomial to address the pitfalls of a Bernoulli or Poisson distribution (Gallagher, 2022). From the source it is inferred that a negative binomial is not restricted to equality between the variance and mean, thus granting us more flexibility. The specification of the negative binomial is directly borrowed from the source and is as follows:

$$\mathbb{P}(\mathrm{K}) = \frac{(K+r-1)!}{(K!\,(r-1)!)} p^r (1-p)^K \quad (4)$$

The letter 'r' describes the number of failures, before 'K' successes with the probability of success 'p.' For further context, the Negative Binomial is used in many fields to model over dispersed count data, where the variance is greater than the mean. Its purpose is thus suitable for spare parts demand because we have learned earlier from Diamond (2019) about intermittent that the local PDF is substantially greater than PDF's mean, i.e. $x \gg \bar{x}$. Moreover, the distribution assumes that the number of successes is random, and the probability of success is the same in each trial. Gallagher encouraged in his presentation to explore the alternative distribution assumptions by putting the negative binomial to further use. In the subsequent year, Kerzel (2023) modelled customer demand as a *negative binomial distribution* **(NBD**). In his report, he argues that the NBD will describe data better than a Poisson distribution as the negative binomial distribution is over-dispersed.[6] Moreover, it is inferred from its report that NBD classifies data well.

2.3.3. Contextual Forecasting

Contextual forecasting addresses the human factor in supply chain forecasting. Pinçe et al. (2021) describe it as an approach to overcome the challenges of spare parts demand patterns influenced by external factors, through systematically combining available information to improve forecasting performance. The authors further explain that contextual forecasting can be divided into judgmental and installed base forecasting.

Judgmental forecasting entails statistical adjustments made by professionals with expertise in operations and supply chain management (Arvan et al., 2019; Goodwin et al., 2018; Perera et al., 2019). Pinçe et al. (2021) assert that statistical forecasts are rarely used without expert input. Their literature review reveals mixed evidence on whether judgmental adjustments improve or worsen statistical forecasts, with some academics suggesting that human intervention can lead to poorer performance due to the lack of quantifiable information.[7] Installed base forecasting, on the other hand, involves using

[6] With over-dispersed the author was implying $\sigma^2 > \mu$.

[7] See Sander (2003) and Franses and Legerstee (2010).

information about equity, quantity, location, failure rate, and other factors to predict spare parts demand (Pinçe et al., 2021).

Even without professional insights, this paper explores the potential of contextual forecasting through a consensus approach based on industry standards. These standards comprise generally accepted rules, specifications, or guidelines followed by firms within a particular industry. To ensure accurate and reliable forecasting results, this paper can be consulted through publications by established organizations that develop and promote industry standards, such as ISO, ISM, APICS or industry-specific aftermarket associations. These organizations offer valuable resources, certifications, and industry benchmarks that can guide the forecasting process when specific data points or professional insights are lacking.

For instance, Syntetos et al. (2005) proposed a taxonomy based on demand intermittency and variability, enabling companies to classify demand patterns and select appropriate forecasting techniques. This approach aligns with the broader concept of adhering to industry standards and best practices in spare parts demand and supply forecasting, as it emphasizes understanding the unique characteristics of demand patterns and selecting suitable forecasting methods. By applying the taxonomy proposed by Syntetos et al. (2005) and adapting forecasting techniques to specific demand patterns, companies can optimize inventory management and enhance customer satisfaction. This approach supports the overall goal of adhering to industry standards and best practices in spare parts demand and supply forecasting.

The objective of the thesis is to integrate industry standards from relevant organizations into the contextual forecasting model. By doing so, the thesis can estimate variables like lead time or average costs, even when the data is unavailable. This approach helps to facilitate informed decision-making and enhance the accuracy of the forecasting process. Additionally, the utilization of industry standards aligns the forecasting methods with industry best practices, enabling organizations to benchmark their performance, identify areas for improvement, and refine their forecasting techniques. It is worth noting that incorporating industry standards may pose challenges since some consensus (proxies) may not be applicable to a specific industry, leading to a stark difference in data.

2.3.4. Data Aggregation

The intuition behind data aggregation stems from the idea to aggregate data with similar demand patterns. According to Pinçe et al. (2021), data may be temporally aggregated or across time series to reduce the number of zero-demand periods and to make more accurate forecasts. From Hyndman and Athanasopoulos (2018) it is inferred that temporal aggregation describes the technique to identify trends and seasonal components of a time series by aggregating the data inputs over a longer period, such as quarters or months. It is also inferred that the potential merits of aggregation are reducing noise in the data and making it easier to identify the underlying patterns and relationships. Simultaneously, temporal

aggregation may result in the loss of information and detail. An additional concern related to temporal aggregation entails choosing the appropriate level of aggregation, which is based on the characteristics of the data. To address the issue, Nikolopoulos et al. (2011) suggested rectifying the problem by setting the aggregation level equal to the lead time length along with one additional review period. That is because the authors denote that period reviews in inventory systems forecasts are computed to determine the safety stocks for the current period.[8]

Aggregation across time series, on the other hand, describes the process of combining multiple time series datasets, which according to Hyndman and Athanasopoulos (2018), can be of the same or different frequencies, into a single time series dataset. The intuition behind the process is to compare or analyze the relationships between different variables or when combining data from different sources to create a more comprehensive dataset.[9] Another benefit is improved accuracy and reliability of forecasts along with noise reduction that helps smooth out noise or short-term fluctuations that may be present in individual time series (Clemen, 1989; Brockwell and Davis, 2016). On the contrary, aggregating across time series may have the potential drawback that leads to an increase in the complexity of the analysis, as relationships between variables in the combined dataset may be more difficult to interpret (De Gooijer and Hyndman, 2006). Additionally, concerns entail the loss of granularity, thus describing the loss of detail in the final aggregated dataset, which potentially obscures important features or patterns present in individual time series (Hyndman and Athanasopoulos, 2018).

This thesis aims to apply data aggregation strategies that diverge between temporal aggregation and across time-series predictions. From existing literature, it is inferred that they are based on traditional time series prediction methods. In the sections to come, the standalone prediction models are highlighted along with their aggregated iteration.

2.3.5. Combined Strategies and The Concern over Complexity

The three highlighted strategies can also be combined to yield more accurate forecasts. From Petropoulos and Kourentzes (2017) it is inferred that standalone forecasts are used at different temporal aggregation levels. Another approach entails combining forecasts from transformed frequencies from the same or multiple methods. It is lastly inferred from the authors that an approach with multiple steps leads to better forecasting results. These results can further be improved by using demand classification schemes.[10] This thesis, thus, wants to contribute to the existing literature by interchangeably using a mixture of combined- and standalone forecasts as well as data aggregation and demand classification strategies to yield better forecasting results. The caveat of combining approaches is the increasing degree of complexity. From the past three sections, all three approaches share the common drawback that the mixed approaches lead to greater complexity. By further combining strategies, the line of reasoning thus follows that complexity further increases at the expense of accuracy.

[8] In the context of spare parts demand, lead time refers to the amount of time it takes to receive a spare part after an order is placed (Axsäter, 2006).
[9] See also Brockwell and Davis (2016).
[10] See

2.4. Model Selection: The Law of Parsimony

This subsection of the literature review discusses the application of Ockham's Razor in academia, specifically in the context of model selection for forecasting. It highlights the danger of over-parameterized and complex models that generalize poorly and emphasizes the importance of finding the optimal number of parameters to achieve accurate forecasts. The review also discusses various model selection techniques, including the use of the Bayes Factor, Automatic Intermittent Demand Selection, and Cross-Validation. Finally, the review suggests that by using Cross-Validation to evaluate and compare models with different complexities, one can identify the simplest model that performs well on the data, in line with Ockham's Razor's principle of preferring simpler explanations.

2.4.1. Ockham’s Razor in Academia

This subsection aims to provide additional support for the existing literature on the topic. As previously established, the law of parsimony, commonly known as Occam's Razor, suggests that simpler explanations are generally preferable. In the context of forecasting, this principle translates to the identification of a model that performs well on the available data while maintaining simplicity. To achieve this, model selection techniques should aim to identify the optimal number of parameters.

It is important to apply these techniques because more complex models, such as those that use deep learning methods with multiple layers, may fit the data better, but are at risk of becoming over-parameterized and too detailed, which can lead to poor generalization (MacKay, 1992). Therefore, Occam's Razor favors simpler methods over needlessly complex ones.

Previous research on Occam's Razor and forecasting has focused on the investment sector. Bogle (1991) argues that long-term stock market forecasts can be accurately assessed with fewer than five elements, satisfying Occam's Razor due to its simplicity. In the 21st century, Estrada (2007) applied Bogle's approach to contemporary circumstances and found that the "simple method" is surprisingly successful in predicting the returns of twelve international stock markets. While neither study referenced or compared complex methods, the question arises as to the trade-offs of using methods that demand greater computing power versus the marginal gains in accuracy.

A study by Graefe et al. (2014) addressed this question and concluded that simple approaches yield better results than their complex counterparts. This finding emphasizes the importance of simplicity in forecasting, even when sophisticated methods are available.

Armstrong and Green (2018) proxied Occam's principle through 15 relatively simple evidence-based forecasting methods and reported substantial improvements in accuracy with simpler methods whilst concluding that more complex methods are unsuited for forecasting. Accounting for the prior and current section, Occam's Razor (i.e., Occam Effect) applied to the thesis's context is therefore a relative concept that depends on the interaction between competing model parameters and their contribution toward making accurate forecasts amid added information. The term 'relative' suggests that one can compare two competing models based on their complexity — with which Occam's Razor would favor explaining phenomena with the model of lesser complexity. Hence, the Occam Effect ought to be perceived with slight nuance as competing models may distinguish themselves in terms of complexity yet are still, in absolute terms, intricate.

From Goodwin (2015) it is inferred that two competing simple alternatives would each lead to the same decision. Nevertheless, Ockham's razor is our 'common sense' because any reasonable being ought to use the most pragmatic approach toward solving phenomena (Sivia, 2009, p.92). The following question would be if one's common sense leads to better results. Findings diverge in a direct comparison between simple and complex methods.

Lastly, literature specifically involving Occam's rule and spare parts demand is not heavily prevalent. A study case on Algeria's Bay Port (Spain), however, investigated ML methods to predict freight congestion (Ruiz-Aguilar et. al, 2016). The authors used a panel to compare forecasting models and concluded that increasing complexity yields no evidence of better predictions.

2.4.2. Bayes Factor and Intermittent Demand Model Selection

The work by Fisher (1935) allowed research through the Fisher matrix approach to estimate parameters by allowing one to predict how well an experiment will be able to estimate the model's parameters. Heavens (2007) expanded on model selection techniques by allowing research to distinguish between different models. Though the author's article's field is focused on the field of Astrophysics, the latter and this thesis share a common interest to elucidate *nested models* that describe more complicated numbers that have a greater number of features. So, the authors proposed a method to distinguish between different models, regardless of their parameters by computing the Bayesian Evidence ratio for two different models. Also known as the ***Bayes factor*****,** the measure intends to quantify the support for one model over its alternative (Romeijn et. al, 2016). But as we have learned in earlier sections about intermittent data, it reiterated that they depart from Gaussian statistics to varying degrees. Methods linked with Bayes Factor are therefore unsuitable in our research because they assume our data to have a Gaussian distribution, which is not the case.

Research in intermittent demand model selection is not plentiful, though the SAS Institute (2020) proposes Automatic Intermittent Demand Selection. The authors proposed the guideline to selected

models considering how well the model predicts the average demand concerning time by using the component prediction errors of the respective time series, seen beneath in the example made by the SAS Institute, the prediction for Croston's method $\frac{\overline{d_i}}{\overline{q_i}}$ and the average demand method $\overline{a_i}$ or both are based on the selected method for each component. The guideline for choosing between either model is as follows: the lower the error prediction the model the better.

Table 1: Prediction Error Components

Definition	Expression
Demand Interval Series	$e_i^q = q_i - \overline{q_i}$ for $i = 2\ to\ N$
Demand Size Series	$e_i^d = d_i - \overline{d_i}$ for $i = 1\ to\ N$
Average Demand Series	$e_i^a = a_i - \overline{a_i}$ for $i = 2\ to\ N$

In academia Kourentzes (2014) also discusses error metrics for Automatic Intermittent Demand Selection by investigating whether it is possible to select between different models for the respective time series, based on the outlined error metrics. Related to complexity, more complex models tend to have lower error criteria and were thus favored in Kourentzes' s paper (Hastie et. al, 2009; James et. al. 2013).[11] Moreover, the author concluded that though demand classification (erratic, lump, smooth or intermittent demand) helps to communicate necessary properties, the model selection result suggested that none of the adjusted methods managed to outperform individual models. In the end, the author emphasizes the need for more valid model selection methodologies.

2.4.3. Non-Parametric Model Selection Techniques and Complexity

When considering model selection techniques, it is important to consider the theoretical properties of model selections. From the previous paragraphs, it is inferred from past research that intermittent data departs to varying degrees from Gaussian statistics, indicating that they are non-parametric because they do follow a known distribution. Ding et.al (2018) created an overview of model selection techniques in which *Cross-Validation* **(CV)** matches the description of a selection method that does not require parametric assumptions. The authors specifically direct the attention to a variant of cross-validation **(CV)** known as the "delete-1 CV" method, synonymous with *leave-one-out* **(LOO)**. The intuition of the approach is as follows: assuming n number of observations, one intentionally leaves out one observation in turn and attempts to predict the observation by using the $n - 1$ number of observations that remains. Subsequently, the average prediction loss over n rounds is recorded. Alternative iterations that aim to

[11] Kourentzes (2014) proposed an automatic model selection that would optimize the paper's forecasting models through a 'KH' item classification based on the different in-sample error of the cost functions sAPIS and MASE.

select the model with the smallest average validation loss are *k*-fold CV and *generalized cross-validation* **(GCV)** which are reported as less computationally heavy.

In later years Zhang et. al (2023) dwelled deeper into the CV method and reported that the latter can be used to select between modelling procedures, such as between traditional modelling and Blackbox approaches. The selection procedures according to the authors are as follows: if the CV approach prefers the traditional approach, it is inferred that in this example liner model, bestows the prediction an accuracy advantage but also with a sensible explanation of the regression relationship. Vice versa, preference toward black-box models, in this case, a neural network model, the linear model may have missed important nonlinear effects.[12] To bridge the gap between CV and complexity, this thesis takes the number of features and parameters of our models into consideration. We thus set the expectation that by using cross-validation to evaluate and compare models with different complexities, one can identify the simpler models that perform well on our data. The latter aligns with *Ockham Razor's* principle of preferring simpler explanations. An indicator is thus the number of features and parameters of our models.

2.5. Real and Synthesized Data Inputs

Synthetic data is becoming more common for its benefits in privacy protection and cost savings. Deep learning methods like GANs and RNNs, specifically LSTMs, have been used for synthesizing data. This thesis focuses on using LSTMs to synthesize spare parts demand data by replicating a recent study that proposed a new architecture using an autoencoder to improve performance. Although there is limited research in this area, the ability of LSTMs to model nonlinear relationships and their success as a discriminative model suggests its potential as a generative model for synthesizing data inputs.

2.5.1. Generating Data Across Various Themes

Synthetic data put simply describes any measurements that are not obtained by direct measurement. The latter ought to approximate real outputs as much as possible as previously mentioned, synthetic data is often used to substitute for erroneous data in contemporary applications. The merits derived from the use of synthetic data are perceived to be privacy protection and cost-saving.

There are several applications aimed at deriving artificial data, but they are mainly designed for specific purposes. One such application within the R-language is called Synthpop. The R package approximates a series of conditional distributions (Nowok, 2017). Synthpop synthesizes data on an identifier-by-identifier basis by fitting regression models and deriving artificial values from the associated predictive distributions. In addition to filling in missing data points, Synthpop generates synthetic data

[12] Zhang et. al (2023) describe the nonlinear effects, such as interactions and higher-order terms.

patterns to account for missing values. Nowok et al. (2017) applied Synthpop to UK longitudinal studies, synthesizing overly sensitive consensus data. The research suggests that a benefit of the application is its ability to quickly automate data, which is crucial for research purposes. The methods applied to the synthetic inputs were deemed sound. However, it is important to note that the synthesized data should only be used as a guide, and the output should not be treated as definitive. The cost-saving aspect was not the focus of the study, and the strengths and weaknesses of using synthesized data are not well established.

Another R-language program used to synthesize data is called *'fabricatr'* (Blair et al., 2022). The goal of the application is to help researchers imagine data before collection by allowing them to easily simulate correlated and hierarchical data structures. The application works by transforming a known classifier into the standard normal space through an affine transformation. The next steps involve computing conditional distributions of the transformed variable as a standard normal and then mapping the standard normal back to the target distribution. The value of the application lies in its ability to model data before analysis, thereby helping to clarify the experimental design of any research. Current literature suggests that synthesizing data corrects issues associated with real data, such as missing entries. However, the limited literature on using synthetic data in real-world scenarios has not fully established its strengths or weaknesses. Nevertheless, it can be inferred from von Bismarck (2022) that artificial inputs can benefit from data-smoothing and augmentation techniques, which correct for periods that are not representative of the data, such as during the COVID-19 pandemic.

Newer and more computationally heavier approaches, to which this thesis dedicates its focus, are found in deep learning methods. One approach developed in recent years is known as the **(GAN)** *Generative Adversarial Network*. Developed by Goodfellow et. al (2020), GAN uses an artificial intelligence algorithm to imitate existing data. Though the application originally intended to synthesize media such as video, images and voice recordings, it can also imitate time-series data. One example includes the paper on solar data by Zhang et. al (2020). A large share of solar data can be found missing, which the authors suggest worsens the data's quality. To solve the issue, the authors utilize GAN on public datasets and reported that the GAN-based data imputation methods minimize the data' mean squared error by at least around 24%. The paper is relevant to spare parts forecasting as the power generation levels of solar panels are intermittent, requiring appropriate forecasting methods and high-quality data.

Examples of challenges in training intermittent data with GAN, according to Saxena and Cao (2021), are mode collapse, instability, and non-convergence, which are caused by improper network architecture or selection of the optimization algorithm. To address these issues, the authors recommend re-engineering GAN's network architectures or using other methods in combination to achieve the desired outcome. One example includes **(IIT-GAN)** *Irregular and Intermittent Time-series Synthesis with Generative Adversarial Networks* proposed by Jeon et. al (2021). As the model's name already suggests, the application ought to synthesize intermittent inputs. The authors aim to achieve the goal by combining an array of deep learning techniques that are beyond this paper's focus. Though ITT-GAN is still under

review, the paper reports that ITT-GAN synthesizes data with superior quality compared to competing methods. However, amid a lack of expert knowledge and computing power, this thesis refrains from exploring using ITT-GAN to synthesize spare parts sales data.

2.5.2. Deep Learning: Recurrent Neural Networks

An alternative to GAN is the *Recurrent Neural Network* **(RNN)**, which is a deep learning method designed to process data sequences. While both GAN and RNN are deep learning models, they differ in their objectives. RNNs aim to process sequential data and capture information about the context of the sequence, while GANs compute new data samples. LSTMs were introduced to solve the issue of vanishing gradients in traditional RNNs. Hochreiter (1997) introduced the LSTM iteration to address the issue of vanishing gradients in traditional RNNs. This iteration prevents the gradients from shrinking during backpropagation over multiple time steps, thereby making the network easier to train. Due to the long-time span of spare parts data, it may be beneficial to apply LSTMs to generate synthetic data. In related research, RNNs are frequently used for natural language processing tasks. Examples include speech recognition, text synthesis and in chat-bots (Sak et al., 2015; Sutskever et. al, 2011; Yin et al., 2017). Behjati et. al (2019), for example, used an LSTM to compute highly accurate data meta-event. From the research, it is inferred that the LSTM was applied on a relatively small data set and encourages future research to apply the latter to larger data.[13]

LSTM use diverges is split into functioning, according to Behjati et. al (2019), as either a *generative or discriminative model*. The goal of a generative model is to capture the joint probability of inputs and outputs, which can then be used to sample data by calculating the most probable output using Bayes rules. Vice versa, the discriminative model captures the posterior probability directly and aims to make predictions in regression and classification tasks. Currently, there is limited research on using Recurrent Neural Networks (RNNs), such as LSTM, to synthesize time series. Related to the subject, research on extreme event forecasting at Uber stated that standard LSTM shows relatively poor performance relative to the state-of-the-art approach (Laptev et al., 2017).[14] Thus, the authors proposed a new architecture, which leverages an *autoencoder* for feature extracting, which allowed them to achieve performances superior to their baseline models. An autoencoder briefly describes a type of artificial neural network which is meant for unsupervised learning (Goodfellow, 2016). Its goal is to learn a compact representation of the input data, coined encoding, and subsequently use the encoding to reconstruct the original data set. Relevant to our research, it is inferred from the paper by Laptev that its RNN iteration can model nonlinear relationships among features and is simultaneously computationally efficient.

[13] In the context of the paper, meta events refer to census data of Norwegians that encompasses demographic categories such as age, gender, race, and education level.

[14] The state-of-the-art approaches are described in the paper as Croston's method and Random Forest.

2.5.2.1. Long Short-Term Memory (LSTM) & Autoencoding Structure:

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed to handle time-series and sequence data by learning long-range dependencies (Hochreiter and Schmidhuber, 1997). The LSTM consists of memory cells with input, output, and forget gates that control the flow of information. These gates allow the network to selectively remember and forget information across time steps, enabling it to capture long-term relationships within the data.

An autoencoder is a type of neural network that learns to reconstruct its input data by first encoding the input into a lower-dimensional representation (also known as a bottleneck or latent space) and then decoding it back to the original input space (Hinton and Salakhutdinov, 2006). The autoencoder consists of two parts: an encoder, which maps the input data to a latent representation, and a decoder, which reconstructs the input from the latent representation. The goal is to minimize the reconstruction error, forcing the autoencoder to learn a compact and useful representation of the data.

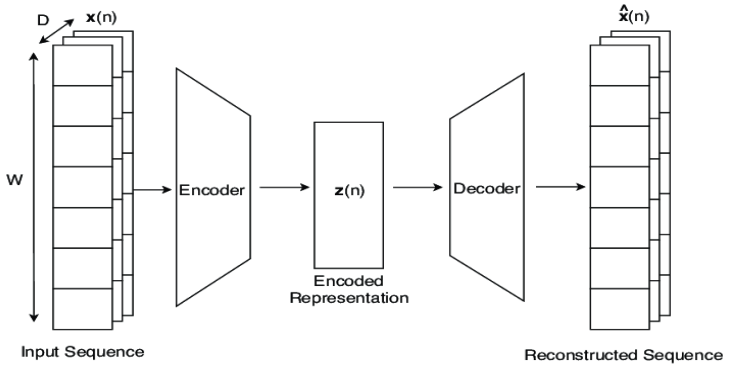


Figure 2: LSTM Autoencoder Structure by Trinh et al. (2019)

An LSTM autoencoder, depicted in Figure 2, combines the LSTM architecture with the autoencoder concept. It is particularly useful for time-series and sequence data, as it can capture temporal dependencies in the data. In an LSTM autoencoder, the encoder is composed of LSTM layers that learn a lower-dimensional representation of the input sequence. The decoder, also composed of LSTM layers, then reconstructs the input sequence from the encoded representation (Chung et al., 2014). By learning to minimize the reconstruction error, the LSTM autoencoder can generate a compact and meaningful representation of the time-series data, which can be used for various tasks such as data synthesis, anomaly detection, and forecasting.

This thesis seeks to replicate Laptev's approach, using the new architecture to evaluate the potential of LSTM as a generative model for synthesizing spare parts demand data. Despite limited research on using RNNs for synthesizing spare parts demand data, the ability of LSTM to backpropagate through the network over many steps and its success as a discriminative model in prediction tasks suggest its potential as a generative model for synthesizing data inputs.

Part II: Literature Review on Methods

2.6. Literature on Spare Part Demand Forecasting

A series of competitions, known as the M-Competitions was motivated in the 1980s to have a benchmark to fairly evaluate and compare the accuracy of time series forecasting methods. Since the M-competitions inception, there have already been 5 competitions — with a 6th competition set to conclude by 2024. For each competition, new methods were introduced to assess intermittent data. The coming subsection further dwells on the M-Competitions along with the different types of methods, classified as non-traditional- and traditional time series — as well as machine learning methods (Baisariyev et. al.,2021). Further classifications are borrowed from et. that entails non-parametric spare part forecasting that splits into data resampling and aggregation approaches as well as deep learning methods (Pinçe et al., 2021).

2.6.1. M Competitions

Amidst a plethora of forecasting methods used to predict the demand for spare parts, it is difficult to establish a consensus on the effectiveness of standalone versus multiple methods through individual comparisons. Instead, deductions are often made using a benchmark known as Makridakis Competitions.[15]

Before discussing the competition, it's briefly explained why methods cannot be individually compared. Forecast measures' reliability is subjected to accuracy metrics that numerically derive a method's ability to predict reliably spare part demand. These accuracy metrics are commonly known as standard forecast error measures, such as the *mean absolute percentage error* (***MAPE*).** The measures are unable to provide fair evaluations for intermittent demand series due to numerous periods of zero demand, hence an accuracy metric may be unable to compute a score because of a zero-value denominator. Thus, Makridakis and Hibon (1979) criticized previous accuracy comparisons – claiming that they were impossible to compute. Consequently, the M-competitions were born.

Makridakis Competitions is described as the first platform with which time-series forecast methods can compete.[16] Though forecasts of competitions within M-competitions' lore are across a broad spectrum; this paper's scope is set on the competitions that assimilate assessments on intermittent demand. The competitions refer to a benchmark involving fixed deductions in the literature that builds upon the M-competitions to evaluate various methods. It has been observed that combinations of primarily statistical

[16] See Makridakis et. al (2020)

approaches yield greater numerical accuracy compared to individual statistical and machine-learning methods (Makridakis et al., 2020). However, the same authors acknowledge that intermittent time series were not included in their evaluation. This exclusion was attributed to the distinct nature of such data (e.g., numerous zero values) and challenges related to accuracy measures for forecasting, which will be discussed in the methodology section.

Table 2: Forecasts in M-competitions Featured in Literature

Method	Competition
Croston's Method	M3
SBA	M3
Bootstrapping	M5
Exponential Smoothing	M3
Neural Networks	M5 & M3
MAPA	M5 & M3
Exponential Smoothing	M5
XGBoost	M5

Note: See Makridakis et al. (1982) and Makridakis (2020)

Subsequent M-competitions were initiated a year after the previously mentioned study, known as the M-5 competitions. These expanded on the M-4 competition by, among other things, also concentrating on time series exhibiting intermittency (Makridakis et al., 2021). Although the authors assert that the M-5 competition surpasses its predecessor, they also admit that the current findings are provisional. This is a reasonable expectation, given that the data inputs are solely focused on retail from the US corporation, Wal-Mart. From the literature, we use the Makridakis Competitions as a benchmark by feeding its forecasting methods with spare parts demand data. The thesis's focus will be on the competitions that assess the accuracy of forecasts for intermittent demand, as well as iterations, combinations and other methods or approaches that may have or have not been used during the competitions to widen our scope. The coming subsection outlines the methods appearing in M-competitions and will also be utilized used in the analysis of this paper.

2.6.2. Traditional Time Series

Standalone

Time series are defined as "*(…) a chronological sequence of observations on a particular variable* (Bowerman et. al, 2005, p.4)." Examples are the volume of purchases over time and the level of inventory. A *Moving Average* **(*MA*)** approach, for example, aims in modeling univariate time series. The MA postulates that the output depends in a linear matter on the current and past values. The perceived benefits of the MA approach are thus its simple mechanism and efficiency in smoothing out short-term

fluctuations, but the drawbacks are that it assumes equal importance of past data points (Hyndman and Athanasopoulos, 2018). Thus, it is deemed ineffective for intermittent demand patterns.

The *weighted moving average* **(*WMA*)** expands on MA by putting more weight on recent data, and less on earlier entries. From the previously mentioned source, the perceived benefit is thus that it allows our forecasts to be more responsive to trends, and thus provides better forecasts for stable data. The drawbacks of this approach, however, are that WMA still struggles with intermittent demand patterns because an optimal weight needs to be determined.

Earlier in our literature review, *Exponential Smoothing* **(*ES*)** was introduced, which is an adjusting technique that accounts for prior periods' forecast and adjusts it either upwards or downwards based on what is occurring (Krajewski et. al, 2012). According to Hyndman and Athanasopoulos (2018), the perceived benefits of ES are that it can capture trends and seasonality and is capable of adapting quickly to changes in demand patterns. On other hand, the perceived drawback is that it requires tuning of smoothing parameters, and thus may not perform well with intermittent data.

Data Aggregation

The three methods in the prior paragraph represent our traditional time series because they appeared frequently in existing literature, albeit proven inappropriate in predicting ID.[17] Combined with different methods, Kourentzes et.al (2014) recently proposed combining the 'Traditional Time Series' methods with a *Multiple Aggregation Prediction Algorithm* **(*MAPA*).** The authors describe the combination in which a process known as '*Aggregation'* takes place, which produces time series like the original whilst excluding non-zero periods. The combined forecasted time series components are subsequently fed to the appropriate methods.[18] Afterwards, the forecasts obtained at the respective aggregation level are subsequently disaggregated back to the original frequency. Finally, a combined forecast takes place to produce a final forecast. This is reportedly done with the previously mentioned traditional prediction methods.

From the paper, it is inferred that the combined approach benefited from improved forecast accuracy. That is because MAPA captures information from multiple aggregation levels, which enables researchers to identify patterns possibly overlooked by our traditional standalone methods. Further benefits include robustness in results because they benefit from the strengths of both aggregate and disaggregate forecasting. Furthermore, MAPA addresses the inability of traditional methods to capture non-linear patterns by detecting them across different aggregation levels. On the other hand, the perceived drawback of the combined approach is the computational complexity of generating forecasts at multiple aggregation levels. Furthermore, guidance in model selection is ambiguous and may require domain knowledge and expertise to ensure that the combination requires better forecast accuracy. Further concerns over the

[17] See Teunter and Duncan (2009).
[18] Kourentzes et. al (2014) reported that the combination led to improved accuracy especially in the long run.

combined approach include determining parameter tuning, thus increasing complexity along with reduced interpretability.

2.6.3. Non-Traditional Time Series

Standalone

Non-traditional Time Series methods were developed to address the shortcomings of traditional time series approaches. In present times, the upcoming methods are frequently used in literature as a benchmark to compare newer methods for predicting intermittent demand. One of which is *Croston's method* **(*CR*)**. Initially proposed by J.D. Croston (1972), he denoted intermittent demand certainly always produces inappropriate stock levels. The explanation for the phenomenon is that demand for constant quantities at fixed intervals can amass stock levels twice the volume needed. Therefore, the author suggested a flexible forecasting model in which separate estimates are made of the probability of demand and occurrence and the size of demand when it occurs. Equation (5) demonstrates Croston's adaptation (Silver et. al, 1998):

$$x_t = y_t z_t \quad (5)$$

Where:

x_t = The demand in period t

y_t = 1 if transactions occur in period t, = 0 otherwise

z_t = Size (magnitude) of transaction in time t

Croston's method assumes that demand is independent between periods, connotating ID. The method also entails an updating process in which only changes are made to an estimate *after* an order has taken place. The benefit of the procedure is that it smooths out the forecast, but infrequent updates possibly introduce a lag in response to magnitude changes (Kourentzes, 2013). It is thus reported that methods like Croston are suited for predicting 'faster intermittent' demand (Syntetos et. al, 2005). The prior findings deflect the method's adhocracy and thus its limits in consistently predicting ID. The latter is nevertheless considered in our analysis as a benchmark relative to other methods. Summarized, CR is a method specifically conceptualized to determine intermittent demand patterns, but its struggles in capturing patterns during periods of low demand. An additional, empirical concern is the need for coherent smoothing parameters.

Another non-traditional method is *Synetos & Boylan Approximation* **(*SBA*)** (Syntetos and Boylan, 2005). The method was proposed in response to correcting Croston's demand estimates mentioned in the prior paragraph. Mathematically, SBA differs from Croston's method because it replaces its smoothing parameter to lessen positive bias with a parameter that intends to lessen the non-zero bias. Since the

proposal, SBA is widely used in other research on intermittent demand.[19] Summarized, this method is improved by Croston's method due to its ability to better handle zero demand periods. Furthermore, a possible drawback of the approach is its assumption of a constant demand rate during non-zero periods.

Later research by Romeijnders et. al (2012) empirically has shown that both Croston and SBA are unable to deal with abrupt declines in intermittent demand, connotating the *obsolescence* of an SKU. In response to the problem, *Teunter-Syntetos-Babai* **(*TSB*)** was proposed by Teunter et.al (2011). The latter distinguishes from Croston's method and SBA by updating the demand probability instead of its interval. The benefit of the approach is that TBA can more quickly react to scenarios of looming obsolescence. Contrary, the authors proposing TBA state that Croston's or SBA would likely not be able to update if items were suddenly obsolete (Teunter et. al, 2011). Hence, TBA distinguishes itself from its alternatives by which it consistently updates through altering probabilities versus an actual purchase taking place. The possible drawbacks of the method, however, are the increased complexity along with the requirement of tuning parameters.

Combined Forecast

Relevant to the recently conceptualized 'analytical pluralism,' the benchmark methods can also be used in conjunction with other methods. Since Kourentzes (2013), researchers further attempted to study segmented forecasting as a type of combined method. Fu et al. (2018) for example proposed an integrated forecasting approach with SBA and a *Recurrent Neural Network* **(*RNN*)** to predict semiconductor product demand. The author's justification for such an approach was to account for one of the neural network's shortcomings of needing large amounts of training data with high computing costs, and it is inferred that the combined approach can lead to better forecasting accuracy.[20] A potential caveat of this paper, however, is the absence of the authors addressing the limitations of their proposed approach, yet a plausible drawback paper is the increased complexity of combining artificial intelligence with traditional forecasting methods.

Literature also dwells on combining the additional method with TSB. For example, Tsao et. al (2019) proposed combining TSB with an *Elastic Net* **(*EN*)** and *Random Forest* **(*RF*)** to predict spare parts demand in the energy sector. From the paper, it is inferred that an approach known as *Hybrid Stacking* **(*HS*)** is proposed which entails combining traditional time series forecasting methods and machine learning methods into a single ensemble. Moreover, HS follows a two-stage process where inputs are processed using TSB and RF separately. Afterwards, the processed inputs are processed using EN to compute the final output. The specification of the 2nd stage is borrowed directly from the paper and is as follows:

[19] See do Rego et. al (2008) and Altay et. al (2008)
[20] The benchmark methods in the paper are as follows: MA, CR, TSB, SBA and RNN.

$$\hat{y}_{HS(via\ EN)} = \beta_0 + \beta_{RF} \cdot \hat{y}_{RF} + \beta_{TSB} \cdot \hat{y}_{TSB} + e \ (6)$$

Where:

β_0: The estimated intercept

β_{RF}: The estimated coefficient for the RF's prediction

β_{TSB}: The estimated coefficient for the TSB's prediction

$\hat{y}_{RF}$: The prediction from RF

$\hat{y}_{TSB}$: The prediction from TSB

From the paper's results, it is inferred that the proposed Hybrid Stacking approach performed best among an array of other methods, including the previously mentioned traditional and non-traditional time series methods. A possible drawback paper of this paper, however, is that it only utilizes one single accuracy metric.[21] The inclusion of more than one accuracy metric may have yielded a different performance. Gutierrez et al. (2008) encouraged for future to combine *Neural Networks* (NN) with traditional time methods. For experimental purposes, an adaptation of the *Hybrid Stacking* approach is proposed to combine NN with CR and, thus subsequently feed it into a simple OLS.[22]

Demand Aggregation

An alternative to methods is demand aggregation approaches which literature extensively focuses on, among, *temporal aggregation*. The latter describes the combination in which a low-frequency time series (e.g., quarterly data) derives from high-frequency time series, such as monthly data (Nikolopoulos et. al, 2011). The drawback, however, is the minimization of historical observations. The same offers thus proposed to apply the before ID, coining it *the aggregate-disaggregate intermittent demand approach* **(*ADIDA*)**. The approach, briefly, consists of three stages. Firstly, determining the type of aggregation, secondly, forecasting the preceding value in the aggregate series, and finally the disaggregation in the forecast into periods equivalent to the original size. It is important to note each stage pertains to various options for the user to decide on, so there is no standard approach in utilizing the prior. From a performance perspective, it is inferred that ADIDA may lead to substantial improvement in a single method's application (Nikolopoulos et. al, 2011). On the other hand, it is inferred from the offers that drawbacks are the computational intensity to make predictions with ADIDA along with selecting an appropriate aggregation level. The following year, Babai et. al (2012) used the latter in conjunction with Croston, SBA, and SES and reported that the hybrid approach is superior to standalone ADIDA. A potential caveat of this combined approach is the increased complexity and determining an appropriate number of aggregation periods. This thesis proposes to use ADIDA in conjunction with SBA, TSB and Corston's to yield improved forecasting results.

[21] The metric used was MASE.
[22] Neural Networks are latter introduced in section 2.6.6. on Deep Learning Methods

2.6.4. Machine Learning Methods

Standalone

Casual methods are a statistical process that is meant to estimate relationships between the dependent variable and the independent identifiers. There are many approaches to assessing the prior, but this paper specifically directs its attention toward a process known as **Gradient Boosting**. The latter describes a machine learning technique used in regression tasks that bestows us with prediction models in the shape of an ensemble of weak prediction models that are often decision trees.

An approach to better conceptualize the method is inspired by Li's (2014) introduction to Gradient Boosting. Briefly, the prior's method purpose is to improve our prediction's accuracy by introducing stage-wise weak learners to compensate for the shortcomings of the already existing weak learner.[23] In the simple example given beneath, we have the prediction F(x1) and actual output value y1:

$$F(x1) = 1.2 \qquad y1 = 1.3\ (7)$$

The subsequent question would be how we can improve our model to the left to attain the actual result to the right — without omitting and or altering anything within our function. The solution that has parallels with gradient boosting is to add a model, typically a regression tree coined *h*. Thus:

$$F(x1) + h(x1) = y1$$

$$F(x2) + h(x2) = y2$$

$$\dots$$

$$F(x_n) + h(x_n) = y_n$$

The simple example above conceptualizes Gradient Boosting, in which in stages the weak learners compensate for the shortcoming of already existing weak learners. Chen and Cuesterin (2016) recently expanded on the concept to develop the process known as *XGBoost*. The latter abbreviated as *Extreme Gradient Boosting* is essentially a scalable, distributed gradient-boosted decision tree, and distinguishes it from its counterpart by building in parallel, unlike Gradient-Boosting's sequential process. Applied to a spare part context, Lama (2020) concluded that current methods in assessing spare parts demand are not ideal, yet *XGBoost* outperforms current forecasting models in certain cases.[24] To summarize it is inferred from the author that XGBoost is a robust and scalable algorithm capable of capturing non-linear relationships and interactions between variables, but requires significant feature engineering and parameter tuning, which can make it more difficult to interpret.

[24] XGBoost would perform better if it would focus on different demand levels for spare parts, which are low ones.

Combined Forecast

Amid the model's recency, literature reporting on methods combined with XGBoost is not numerous, yet Mitra et. al (2022) recently concluded in their research that methods combined with XGBoost outperformed, itself as a standalone. The proposed hybrid is coined ***RF-XGBoost-LR*** and was meant to predict the forecasting of a retail chain. The processes involved in the latter approach involve subjecting inputs to a *random forest* (***RF***), followed by XGBoost, and finally a *linear regression* (***LR***).[25] Briefly, from the authors, it is inferred that the RF ought to make paralleled decision trees to reduce issues involving overfitting, yet they are reported to suffer from minimal training error.[26] XGBoost compensates for RFs shortcomings because it combines multiple weak learners in sequential methods. The predictions are combined in one dataset and finally, a linear regression (LR) processes the final output. The linear regression equation is borrowed and adapted from the paper as follows:

$$Y = \beta + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad (8)$$

Where:

Y are the respective spare parts demand.

β is the intercept.

x are the respective XGB and RF predictions.

From the authors, it is inferred that the hybrid approach improved accuracy due to reduced variance and enhanced robustness to outliers. Some of the limitations of the paper include ambiguity over the required training size along with normalizing the data to make predictions, which may distort interpretations. This paper ought to reproduce the hybrid method with the slight iteration that it will not normalize the dataset. The justification for this decision is to keep predictions uniform to allow for comparison between models without altering their interpretation.

2.6.5. Deep Learning Methods

Standalone

Another alternative in predicting spare part demand is deep learning methods such as the *neural network* (***NN***) is used to forecast intermittent time series, and they ought to provide dynamic demand rate forests that do hold the assumption of constant demand rates in the future (Kourentzes, 2013). Further NN is meant to capture interactions between the inter-arrival rate of demand events and non-zero demand. What makes the prior method lucrative for analysis is the absence of prior knowledge (Dahl and Hylleberg, 2004; Zhang et al., 1998).

[25] Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to make predictions (Breiman and Cutler, 2016).
[26] Each Tree is learned autonomously, thus complementary information from other trees is not accounted for.

A neural network's functional form is borrowed from Kourentzes (2013), as seen below. Intuitively, a neural network can be perceived as a network of *neurons* that are structured in layers. Furthermore, the NN entails predictors on the bottom layer, the forecasts (outputs) on the top layer — and the prior may also have hidden layers.[27] It is said that the absence of hidden layers approximates the alike of linear regressions.

$$Y_t' = \beta_0 + \sum_{h=1}^{H} \beta_h g\left(\gamma 0i + \sum_{i=1}^{I} \gamma\, hiPi\right) \quad (9)$$

The inclusion of the prior adds further complexity which may proxy hidden patterns in intermittent/spare part demand. Referring to equation 3, $\beta[\beta..\beta H]$ are the network weights and $[y_{11}, \ldots y_{HI}]$ resembles output and hidden layers respectively. Further, from the author's equation, it is inferred that $\gamma 0i$ and β_0 are the biases of the earlier-mentioned neurons, which proxy the intercept in a regression for each neuron. $g(\ldots)$ is the non-linear transfer function that provides the non-linear capabilities to the model. H is the number of hidden nodes. From a performance perspective, it is inferred that NN outperforms, among others, the previously mentioned SES (Rumelhart et. al, 1998). On other hand, it is noted that NN's major drawback is its *data hungriness* the demands large samples to train on (Kourentzes, 2013, p. 3).

Combined Forecast

Literature on spare parts further reports that neural networks can also be used in conjunction with different methods. One example includes the research by Dai and Hai (2017) that uses neural networks in conjunction with a support vector machine ***SVM***.[28] The combined method's intuition is, firstly, that SVM and NN compute two separate forecasting models, and finally, the NN model establishes a nonlinear combination forecasting model, based on the two single forecasting models. The motivation behind the final step is that non-linear predictions consolidate with the needs of actual operating systems dealing with spare parts. The authors report that the combined approach yielded higher accuracy in demand forecasts. The findings, nevertheless, need to be perceived with a note of caution as caveats are not explored with the exception that considerable knowledge is needed in the respective industry in which spare parts are offered. Further literature that combines neural networks with classification methods is mentioned by Lolli et. al (2017) that combines the latter method with an *extreme learning machine* to predict automotive dataset. From the authors' findings, it is inferred that evaluation metrics yield evidence for superior performance, but differences in bias remain the same.

[27] See Hyndman and Athanasopoulos (2018)
[28] SVM describe supervised learning models with associated learning algorithms that assess data for classification as well as regression analysis (Cortes and Vapnik 1995).

2.6.6. Resampling Techniques

Standalone

Another alternative to predicting spare parts was proposed by Willemain et al. (2004). It is known as **Bootstrapping** (BS) which uses a two-stage Markov chain to compute non-zero demand points, and then resample the demand using historical inputs. Intuitively, the method avoids making forecasts with values that were previously the same, coined a jittering process enabling more variation. The aspect is that the method preemptively forecasts a series of zero and non-zero demands in the hope to see the next step depending on the transition matrix probabilities. This method is said to be significantly superior to Croston's method and SES at forecasting lead-time demand (Waller, 2015). BS's further merits include the absent need for distributional assumptions, such as data size and behavior (Mobarakeh et. al, 2017). Thus, BS's implications are thus useful in predicting spare parts demand. Nevertheless, the method's caveat is that it is deemed ad-hoc. Further, it is inferred from Babai et. al (2020) that the latter has proven to be effective for intermittent prediction enhancement.

Combined/Data Aggregation Forecast

The BS method's effectiveness can be further improved by incorporating machine learning techniques and other approaches, such as the wavelet-bootstrap-ANN (WBANN) proposed by Tiwari and Chatterjee (2010). Although their method is ad hoc, it demonstrates the potential for reliable hourly flood forecasting. It is suggested that excluding the BS method leads to less accurate results, regardless of whether the wavelet and ANN components are used together or separately. In summary, the wavelet analysis component decomposes the original time series into different frequency components, capturing localized variations and trends in the data. Meanwhile, the bootstrap component enhances the model's performance by generating new samples from the original dataset. Lastly, an Artificial Neural Network is employed to detect complex patterns and relationships in large datasets.

However, the main drawback of the WBANN approach is its increased complexity. Combining three methodologies makes interpretation more challenging and necessitates additional computational resources, which may compromise efficiency. Additionally, the study does not provide a comparison with other existing methods, making it difficult to assess the true merits of the WBANN approach. Despite these challenges, the method's potential for accurate forecasting in various domains should not be overlooked.

2.6.7. Others: Combined Strategies

This subsection addresses an approach by Fu and Chien (2019) called the **UNISON** data-driven intermittent demand forecast framework, designed to enhance supply chain resilience, particularly in the electronics distribution industry. The framework combines traditional methods, such as Syntetos-Boylan *Approximation* (**SBA**) and *Autoregressive Integrated Moving Average* (**ARIMA**), with advanced deep

learning techniques like *Recurrent Neural Networks* (**RNNs**) and employs temporal aggregation to improve the accuracy of intermittent demand forecasts.

The UNISON framework is structured into several steps, depicted above in Figure 3. First, the original demand is transformed into multiple time series at various aggregation levels to reduce the impact of temporal effects on demand uncertainty. At each sub-aggregated forecast stage, three demand forecast models with distinct characteristics are utilized as decision strategies. These models include the Syntetos-Boylan approximation, ARIMA, and RNNs. A weighted combination schema based on decision regret is adopted for forecast combination at sub-aggregated levels. Finally, the forecast results at different aggregation levels are integrated using the disaggregation procedure to produce a unique forecast result.

The methodology adopted in the UNISON framework offers several benefits. By combining traditional, non-traditional, and deep learning techniques, the framework can capture various aspects of the data, leading to a more accurate and robust forecast. Additionally, the use of temporal aggregation allows for a more comprehensive analysis of demand patterns, further improving the accuracy of the forecasts. The method of this empirical study was originally conducted in a semiconductor distributor for validation. It is inferred from the paper that the UNISON framework had better performance than conventional approaches. However, to better understand the improvement, specific performance metrics in the upcoming section will allow for a fair comparison between the rest of the previously mentioned methods.

Considering the inherent complexity of the UNISON framework, it is essential to discuss the scalability and computational efficiency of the approach, especially when applied to large datasets or real-time forecasting tasks. The choice of parameters in the UNISON framework can significantly impact its performance. Therefore, it is crucial to provide information on the sensitivity of the framework to these parameters, such as the number of aggregation levels, the choice of weighting for the combination schema, and the parameters of individual forecasting models (SBA, ARIMA, and RNNs). A discussion on the methodology for selecting or tuning these parameters would be valuable.

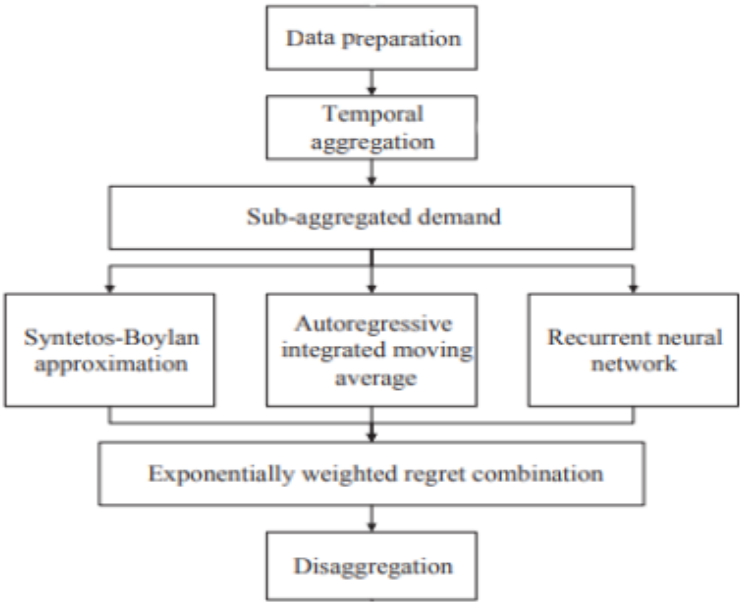


Figure 3 Theoretical Framework by Fu and Chien (2019)

Although the framework was tested in the electronics distribution industry, discussing its potential applicability and transferability to other industries with intermittent demand patterns would be beneficial. Finally, further elaboration on the limitations of the UNISON framework and potential future research directions to address these limitations or enhance the forecasting methodology would contribute to a more comprehensive understanding of the approach.

3. Experimental Design & Methodology

The methodology section outlines the use of various prediction methods to forecast spare parts demand, categorized into Traditional Time Series, Non-Traditional Time Series, Casual Methods, and Others. Forecasting results are evaluated using accuracy and stock-keeping metrics to understand the model's performance. Counterfactual analysis is conducted to assess differences between real and synthetic inputs, while model selection is achieved through Cross-Validation (CV) methods. The study aims to balance fill rate and holding cost in spare parts inventory optimization and adhere to Ockham's Razor principle by selecting the simplest yet effective model.

3.1. Experimental Design & Environment

The experimental design depicted in Figure 4 is based on the literature review of the previous section. The initial stage of the experimental design is to subject our inputs to certain preprocesses, including processing and synthesizing our data appropriate for the analysis. The subsequent step involves predicting

our spare parts sales data via an array of strategies. To reiterate from the literature review, the prediction approach is split into standalone- and combined forecasts as well as data aggregation and demand classification. Furthermore, the results of the forecasts are used to determine which model performs better on the given dataset to which a demand classification approach is applied to match the better-performing methods.

The predictions are evaluated by using the performance metrics to determine the predictive and stock-keeping performance of our models. Subsequently, the model selection ought to determine whether *Ockham's Razor* holds, and finally, the counterfactual analysis to isolate the impact of synthetic- and real data. The coming subsections will thoroughly highlight the evaluation tests used during the evaluation. The analyses for this thesis are made with the Python language. The code for the tests will be made publicly available in the form of five distinctive Jupiter Notebooks for the respective dataset, which will be made available on GitHub. The experimental environment for this thesis is as follows:

- NVIDIA® GeForce® RTX-Series Graphics Card
- 10th Gen Intel® Core™ 2.3GHz Processor
- 32GB Dual-Channel RAM

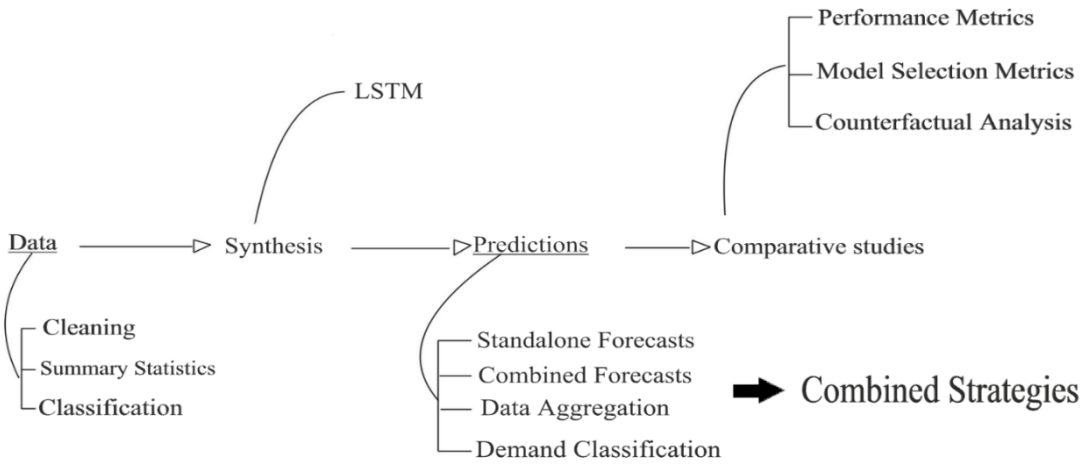


Figure 4: Experimental Design

It is noted that this study involves splitting the data into training and test sets using a cutoff date that creates a 70/30 distribution between the two sets. This split is based on the chronological order of the data, where the data is first sorted by date and then separated into training and test sets. Specifically, all data points that occur before the cutoff date are assigned to the training set, while all data points on and after the cutoff date are assigned to the test set. This approach emulates real-world scenarios, where a

model is trained on historical data and used to make predictions on new, unseen data. Additionally, the split enables us to evaluate the model's performance on unseen data, which provides an assessment of its ability to generalize to new scenarios. Furthermore, the value of smoothing is important for an array of methods in our analysis. In this study, the constant value alpha is 0.15 based on other related studies (Kaya et al. 2020; Boylan and Shale 2003; Teuntner and Duncan 2009). The number K-folds for the cross-validation tests are kept at a default of 10.

3.2. Methodology

Table 2 in the Appendix highlights the standalone and hybrid methods mentioned in the literature review to predict spare parts demand o. These also are the methods with which the paper's research questions will be answered. The methods are classified into four different categories: Traditional Time Series, Non-Traditional Time Series, Casual Methods, and Others. Within each category, there is also a hybrid iteration. The method's implementations are replicated from past research to ensure adequate execution. The coming subsection further discusses how the forecasting results are evaluated.

3.3. Evaluation: Performance Metrics

3.3.1. Accuracy Metrics

Literature on spare parts demands forecasting uses two distinct types of performance metrics: forecasting accuracy- and stock control metrics. The accuracy metric grants us insight into the deviation between predicted and realized demand, while stock control metrics inform us about our forecasting model's implications. Though the prior does not relate to forecasting spare part demand, past literature yields evidence that stock control metrics' inclusion allows for better supply chain management (Dekker et. al, 2013). Further, the average between stock control and accuracy measure suggest they are inherently distinctive, indicating both measures grant two completely different insights. For example, research has shown that methods can distinguish from one another by having either higher forecasting accuracy and service level and/or lower stock level respectively (Syntetos et. al, 2005; Kourentzes, 2013). Combined it is more than plausible to derive that cojoining both performance metrics allows researchers to better understand phenomena – especially when inferring from Walström and Segerstedt (2010) that one performance measure alone is unable to represent all forecast error's dimensions.

In earlier sections, frequently appearing accuracy metrics, such as MAPE were discussed along with their inability to reliably evaluate a method's performance in predicting spare part demand amid many zero-value periods. Hyndman (2006) found a way around the issue by comparing accuracy measures for the time series forecast used in the M-competitions. They found that the performance used is subjected to issues that give infinite or undefined accuracy scores when attempting to predict intermittent demand. To go around the issue, the authors suggest that the forecast errors need to be scaled by the in-sample absolute error, computed by using the naïve forecasting method. Hence, the same authors proposed ***MASE***, which is widely applicable due to the measure's ease of interpretation; ***MASE*** values greater than

one indicate tendencies for worse forecasting performance than in-sample forecasts from the naïve method and vice versa. The measure is computed as follows:

$$MASE = \frac{1}{n}\sum_{i=i}^{n} |\, q_t| \quad (10)$$

Where q_t are the scaled error and its determination is derived as follows:[29]

$$|q_t| = \begin{cases} \dfrac{|e_t|}{\frac{1}{n-1} * \sum_{t=2}^{n} |y_i - y_{i-1}|} \\ \dfrac{|e_t|}{\frac{1}{n-m} * \sum_{t=M+1}^{n} |y_i - y_{i-M}|} \end{cases} \quad (11)$$

Where:

$\{y_i\}$ is the actual observation time series.
$\{e_i\}$ is the forecast error for a given period.

The two equations above diverge between non-seasonal and seasonal demand on the top and bottom respectively. Where $\{y_i\}$ are actual observations and $\{e_t\}$ the forecasted error for a given period. The ***MASE*** approach is said to be used for comparing different forecasting methods, and the intuition follows that the lower the value, the lower the relative absolute forecast error, and the better the method. The prior method, nevertheless, may be at risk of computing undefined measures because it allows divisions by zero, which is why the decision is motivated to use *Root Mean Square Error* ***RMSE*** as an additional accuracy metric (Martin et. al, 2020, p.3). The latter is an absolute measure and suited for general-purpose error metrics and its interpretation is borrowed from Christie et. al (2022):

$$RMSE = \sqrt{\sum_{i=1}^{n} (S_i - O_i)^2} \quad (12)$$

S_i are the variables predicted values and n is the number of observations given in our analysis. O_i are the number of observations available. It is noted that this metric is suited to compare the model's forecasting errors of a particular variable.

[29] See Hyndman and Koehler (2006)

A modification of Hyndman and Koehler's (2006) Mean Absolute Scaled Error (MASE) is *the Root Mean Squared Scaled Error* **(*RMSSE*)** and addresses the issue related to MAPE that results in an uneven overall error depending on the over- or underestimation of the model because it uses the predicted- and actual values of the test set to scale the MAE.[30] The specification of RMSSE is borrowed from the M5-Competition guide ad seen below. From the latter source, it is inferred that it solves the issue by avoiding using the training data when scaling the Mean Squared Error (MAE).

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \hat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}} \quad (13)$$

Where:

$\hat{Y}$ is the value predicted by the respective model.

Y_t is the actual value to be predicted.

n is the size of the training dataset.

h is the size of the test dataset.

The last accuracy metric is an iteration of the previously mentioned MAPE, known as *Systematic mean absolute percentage error* (***sMAPE***). The latter is synonymously known as "adjusted MAPE" and was proposed and modified in different variants to address the shortcomings of its counterpart, MAPE (Flores, 1986; Armstrong, 1985). To iterate the shortcomings entailed the impossibility for MAPE to handle the occurrence of zero periods of demand along the bias of more heavily penalizing positive errors than negative errors (Hyndman, 2006). From the sources, it is inferred that SMAPE is an accuracy metric based on relative (or percentage errors) and distinguished from MAPE by having both a lower- and upper-bound. sMAPE is thus conceptualized as follows:[31]

$$sMAPE = mean\left(100\frac{|Y_t - F_t|}{Y_t - F_t}\right) \quad (14)$$

Where:

Yt: actual value at time t

Ft: forecast value at time t

|Yt − Ft|: absolute error at time t

(Yt + Ft): sum of actual and forecast values at time t

[30] See Pseudo Lab (2020)
[31] See Makridakis (1993)

100: constant scaling factor for percentage error calculation

mean: function used to calculate the average value of sMAPE across all periods

3.3.2. Stock Keeping Metrics

In this section, the relationship between accuracy measures and optimal inventory performance in spare parts demand prediction is examined. The trade-off curves, fill rate and holding cost, are used as metrics to assess stock-keeping performance. The section also acknowledges data limitations and the necessity of approximations, such as industry-consensus proxies. The *Stock-keeping-oriented Prediction Error Costs* (**SPEC**) metric is introduced as a solution when precise data is unavailable, as it offers a comprehensive assessment of forecast errors by accounting for opportunity costs, thus aiding in more informed inventory management decisions.

Addressing Trade-off between Fille Rate & Holding Costs:

In the realm of spare parts demand prediction, accuracy measures are commonly employed, but they may not necessarily guarantee optimal inventory performance (Syntetos and Boylan, 2006; Syntetos et al., 2010; Teunter and Duncan, 2009). The fill rate, sometimes referred to as the trade curve, is a widely adopted inventory performance measure. Another prevalent metric is the tradeoff curve, which illustrates the equilibrium between service levels and inventory costs. However, according to Pinçe's (2021) review of spare parts demand forecasting research, only 1 out of 29 papers incorporated inventory performance measures.

To measure stock-keeping performance, one can use the service level and tradeoff curve as inventory control metrics since they are the most widely used accuracy measures in the field. The tradeoff curve typically represents the trade-off between the cost of holding excess inventory and the cost of not having the necessary stock to meet demand. Alternatively, stock keeping performances can be measured with the trade-off between fill rate and holding cost (Pinçe et al., 2021; Van Wingerade et al., 2014). The holding cost is calculated as a percentage of the total inventory value or cost. Specifically, it is calculated using the following formula:

$$\textbf{Holding cost} = \textbf{Average inventory level x Holding cost percentage}$$

Where:

$$\textbf{Average inventory level} = \frac{(\textbf{Opening inventory} + \textbf{Closing inventory})}{\textbf{2}}$$

$$\textbf{Holding cost percentage} = \textbf{the cost of carrying one unit of inventory}$$

On the other hand, the fill rate is calculated using the following formula:

$$\textbf{Fill rate}\ (\%) = \left(\frac{\textbf{Number of units shipped from available inventory}}{\textbf{Total number of units ordered or demanded by customers}}\right) \textbf{x 100}$$

The evaluation of the inventory management capabilities of the predictive model involves several variables, such as holding cost and lead times, which are crucial for stock-keeping tests. However, given datasets may not provide all these variables, so a way around is to make use of industry consensus proxies to approximate them. Using holding costs between 20% and 30% of the total inventory is a reasonable assumption and taking a holding cost of 25% per SKU offers a middle-ground approximation (McCue, 2021). This estimation may still not perfectly represent the actual holding costs across industries and businesses, but it serves as a practical starting point. Approximating lead time by assuming a constant lead time for all materials simplifies the calculations, but it is a significant limitation. The lead time can vary depending on the material, supplier, and transportation circumstances. This assumption might not accurately reflect the complexities and variability of lead times in real-world scenarios.[32]

When precise data is unavailable, such as in the case of this paper, new methods like the Stock-keeping-oriented Prediction Error Costs (SPEC) metric can be employed to account for hypothetical losses due to missing inventory. Proposed by Martin et al. (2020), SPEC penalizes forecasts that fail to predict necessary stock levels, considering opportunity costs from demand underestimation or overestimation. This approach offers a comprehensive assessment of forecast errors, enabling better inventory management decisions and thus highlighted in the upcoming subsection.

Stock-keeping-oriented Prediction Error Costs (SPEC)

To contribute to existing literature, a new stock-keeping metric takes the spotlight to account for theoretical costs, known as opportunity costs. It is inferred from Martin et al. (2020) that there are theoretically incurred costs if a spare part is unavailable over a time horizon. The latter is synonymous with commonly known opportunity costs, and the mentioned authors think forecasts ought to be penalized more severely if forecasts fail to predict the necessary stock that suppliers ought to keep in anticipation of future purchases. Thus, the *Stock-keeping-oriented Prediction Error Costs* ***(SPEC)*** metric was proposed and conceptualized on the following page.[33]

[32] The granularity of the available datasets is monthly and weekly data. The lead time for monthly data is kept at a constant of 3 months, and for weekly data 12 weeks.
[33] SPEC metric is currently available within the Python-language.

$$SPEC_{\alpha_1,\alpha_2} = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=0}^{n}\left(max[0; min\left[y_i; \sum_{k=1}^{i} y_k - \sum_{j=1}^{t} f_j\right] \cdot \alpha_1;\ \min\left[f_i; \sum_{k=1}^{i} f_k - \sum_{j=1}^{t} y_j\right] \cdot \alpha_2\right] \cdot (t - i + 1) \quad (15)$$

Where:

n: the total number of periods.

t: the current period.

i: the current period within the sum

y_i: the actual demand at the time i

f_i: the forecast demand at the time i

α_1: the weight assigned to stock-keeping costs

α_2: the weight assigned to lost sales costs.

Equation 15 is directly borrowed from the authors and suggests as follows. SPEC's underscores α_1 and α_2 are the parameters opportunity and stock-keeping costs respectively where both $\in [0, \infty]$. Further, the authors ensure numerical comparability amid changing cost ratios by having a sum of 1 with both parameters combined. The time series' length is labelled by n, whereas actual demand is labelled at time t characterized by y_t. What SPEC does is it calculates each time step's error time whilst penalizing every hypothetical Stock Keeping Unit (SKU) gap at the current time step. Any over- or underestimation is thus say creating opportunity costs. It lastly noted by the authors that the suggested weight distribution between stock-keeping costs and lost sales is 75/25.

Another motivation to use this stock-keeping metric is because SPEC differs from the traditional trade-off curve by specifically accounting for opportunity costs associated with underestimating or overestimating demand. It provides a more comprehensive assessment of the true impact of forecast errors on inventory management decisions, offering additional insight into the effectiveness of demand forecasting models and aiding businesses in making more informed choices regarding their inventory strategies.

3.4. Evaluation: Counterfactual Analysis

Counterfactual analysis is a critical component of impact evaluations and is primarily used to assess the changes attributable to a specific intervention, as well as any unintended consequences. Counterfactuals can be intuitively understood through the following example: Given X, the outcome Y would have occurred, and without X, we get outcome Y'. As Dandl and Molnar (2020) suggest, counterfactual analysis requires imagining a hypothetical reality that contradicts observed facts. In

essence, the analysis aims to identify numerical differences between inputs and methods by considering their respective counterfactuals.

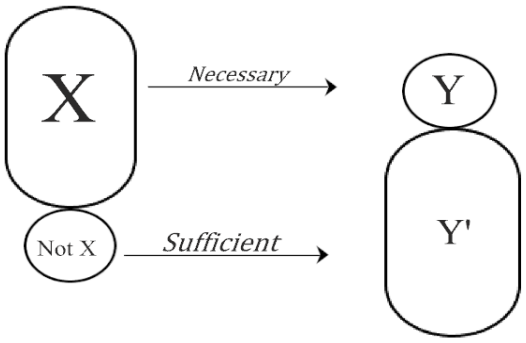


Figure 4: Counterfactual analysis's visualization inspired by Mahoney and Barrenchea (2017).

This thesis takes interest in the differences between real and synthetic inputs in spare parts forecasting models. To assess the differences in accuracy metrics, a counterfactual analysis is used in conjunction with an Out-of-Sample Error (OOSE) approach. The latter describes a statistical method used to evaluate a model's accuracy by testing its performance on data that was not used to fit the model (James et. al, 2013). The OOSE approach involves splitting the data into two parts: a training set with which we train our model, and a testing set to evaluate the accuracy of the model. The model's performance is evaluated based on the error between the actual and predicted values in the testing — which is the core counterfactual analysis.

To determine if the differences between the **OOSE** of real and synthetic inputs are significant, a statistical hypothesis test, such as a paired t-test or Wilcoxon signed-rank test, can be employed. One of the key accuracy metrics for this analysis is the Mean Absolute Scaled Error (MASE), a unit-independent measure that allows for comparisons across different time series. MASE is particularly useful in this context because it facilitates a direct comparison of the forecasting performance when using real and synthetic sales data.

The hypotheses are:

H0: *No significant difference between the OOSE of real and synthetic inputs.*

H1: *Significant difference between the OOSE of real and synthetic inputs.*

A chosen test is performed on MASE and other metrics with an appropriate significance level (e.g., $\alpha = 0.05$). If the p-value is lower than α, the null hypothesis is rejected, indicating a significant difference for that metric. Effect size measures, like Cohen's d, further quantify the practical significance of these

differences. A small effect size suggests limited practical implications, while a large effect size indicates substantial implications for inventory management and cost optimization in the spare parts industry.

In summary, combining OOSE, counterfactual analysis, statistical hypothesis tests, and effect size measures provides a comprehensive evaluation of model performance using real and synthetic sales data, crucial for understanding the impact on forecasting models and the spare parts forecasting.

3.5. Evaluation: Model Selection Through Cross-Validation

In this study, we focus on model selection techniques, specifically Cross-Validation (CV) methods, which are well-suited for non-parametric intermittent data due to the absence of parametric assumptions. Ding et al. (2018) discuss the "delete-1 CV" or leave-one-out (LOO) approach, in which one observation is excluded in turn and predicted using the remaining observations. Other less computationally intensive alternatives include k-fold CV and generalized cross-validation (GCV). Zhang et al. (2023) further explore CV methods, demonstrating their effectiveness in choosing between traditional and black-box modeling approaches. CV's preference for either traditional or black-box models, such as neural networks, provides insights into the presence of linear or nonlinear effects in the data. Initially, this thesis aimed to investigate the relationship between CV and model complexity by considering the number of features and parameters in each model, but ultimately decided to assess the causality between overall complexity and CV, depicted below in **Table 4A**.

Table 4A: *Model Complexity Rankings*

Complexity Tier	Method(s)
Low (5)	Moving Average (**MA**), Weighted Moving Average (**WMA**) Exponential Smoothing **(ES**) Bootstrapping (**BS)** Croston's (**CR**),
Moderate (4)	Syntetos & Boylan Approximation (**SBA**) Teunter-Syntetos-Babai (**TSB**), Extreme Gradient Boosting (**XGBoost**) Multiple Aggregation Prediction Algorithm (**MAPA**)
High (3)	Neural Network (**NN**) Recurrent Neural Network (**RNN**) (**SBA)** & **(RNN)** Multiple Aggregation Prediction Algorithm (**MAPA**)
Very High (2)	Elastic Net (**EN**) & Random Forest (**RF**) Neural Network (**NN**) & Linear Regression (**LR**) Random Forest (**RF**) & Linear Regression (**LR**) Support Vector Machine (**SVM**) & Neural Network (**NN)**
Extreme (1)	UNISON (**RNN**, **SBA** & **ARIMA**) ADIDA (**CR**, **TSB** or **SBA**)

Note*: A more extensive explanation for each categorization is found in Appendix A under* ***Table 4 B***

To identify a correlation between the complexity tiers and the k-fold CV loss, we will use cross-validation to evaluate and compare models of varying complexities. Based on our literature review, we have categorized the forecasting methods into five tiers of complexity: Low, Moderate, High, Very High, and Extreme, as seen in Table 4 above. Our hypothesis is that simpler models with lower complexity tiers will perform comparably to more complex models, thus adhering to Ockham's Razor principle.

This paper will analyze the k-fold CV loss in relation to the complexity tiers to determine if there is a discernible correlation that can guide model selection and improve forecasting accuracy. To do this, we will calculate the average performance of each method across the test datasets, normalize the performance scores, calculate the complexity score for each method, and create a composite score that combines performance and complexity. By ranking the methods based on their composite scores, we can identify the simplest models with the best performance, adhering to Ockham's Razor principle.

Utilizing this approach, we hope to balance the trade-off between model simplicity and predictive performance, contributing to a more robust and efficient forecasting process. If our hypothesis holds, it will indicate that simpler models are preferable in spare parts demand forecasting, provided that their performance is comparable to more complex models. This finding would support the application of Ockham's Razor in this specific domain and help guide future research on model selection and forecasting techniques.

4. Data Description and Classification

This section ought to further elaborate on our data. For this paper analysis, a total of 5 datasets are being used. Four out of the five are modified datasets provided by de Haan (2021) which are publicly available on GitHub. The other dataset also originates from GitHub and is meant for short-term spare parts forecasting. [34] In the coming subsections, the actual data is briefly highlighted along with the measures taken to make the inputs fit for synthesis and further analysis. What follows is the subsection on the data synthesis via LSTM in which steps are described to generate synthetic data. Finally, our inputs are classified based on demand characteristics mentioned in the literature review.

4.2. Data: Structure

The datasets in this research are in panel form, which consists of multiple entities observed over a period. In this case, the entities are the different SKUs (spare parts), and the observations made over various dates. The structure of the data can be described as follows:

[34] See GitHub.

- ***1st Material (int64):*** *A unique identifier for each spare part.*
- ***2nd Price (float64):*** *The price of the spare part.*
- ***3rd Date (datetime64[ns]):*** *The date when the observation was made.*
- ***4th Sales/Synthesized:*** *The actual recorded sale for the spare part at the given date.*

It is noted that this data structure was consciously chosen to increase the degree of freedom in the analysis.

4.1. Data: Description, preprocesses & Summary Statistics

The first dataset coined ***OIL*** is an industrial dataset containing sales data on 1423 spare parts for an undisclosed oil refinery.[35] The data is in monthly format ranging from January 1997 up to August 2001. For the analysis, the dataset ***OIL*** posed two empirical challenges of which one is the presence of negative values and the second the introduction of new stock-keeping units (SKUs) that may distort our analysis results. The first challenge was simply treated by inputting the negative value to 0. The second challenge was approached by assessing the median and mean row entries of the dataset with values greater than 0. [36] It is reported that the median number of entries is lower than the mean, which suggests there were a few periods with unusually high sales. Another interpretation of the difference is that products were introduced later into the dataset, and sales before their introduction were simply recorded as 0. The decision was thus made to omit rows with less or equal to 26 entries. The final number of SKUs in our analysis is thus 9933 spare parts.

The second dataset coined ***AUTO*** and entails sales of 3000 SKUs of a firm operating in the automotive industry. The data was used in the report by Syntetos and Boylan (2005). It is reported from the paper that the data is the sample of an originally greater dataset, which specifically contains SKUs with faster intermittency. Furthermore, the data is in monthly format ranging over an unspecified 2-year period, and all SKUs, according to the authors, are treated as single units as opposed to being sold in packs. During the preprocessing stage, it is noted that this data set raised any noticeable empirical concerns. From de Haan (2021) it is inferred that the dataset lacks lead times and prices. Thus, to proxy prices, this paper decides to follow Zip's (1935) law that describes the relation between rank order and frequency of occurrences and suggests that observation is inversely proportional to its rank. The strong assumption is therefore made that market forces prefer more inexpensive SKUs and is thus reflected in the number of sales.[37]

[35] This dataset was used in the paper by Porras and Dekker (2008).
[36] The dataset's original format was cross sectional and is reshaped as a panel set for further analysis.
[37] The implications of this strong assumption are discussed in the conclusions of this paper.

Table 4: *The Summary Statistics of each Dataset*

Metric	OIL	AUTO	BRAF	MAN	ST
N	546306	68976	419976	517626	1759545
Mean	0.99	4.45	1.44	2.6	2.28
SD	18.17	10.87	16.17	105	51.79
Min	0	0	0	0	0
25 pct.	0	1	0	0	0
50 pct	0	2	0	0	0
75 pct	0	4	0	0	0
Max	3501	416	2062	49980	20000
Var	330.07	118.19	261.64	11027.76	73373.69
Granularity	Monthly	Monthly	Monthly	Weekly	Monthly

Note: The continuous values are rounded to the 2nd decimal.

The third dataset coined ***BRAF*** is from the Royal Air Force (UK), entailing sales of 5000 spare parts for aircraft over the period 1996-2002. From de Haan (2021) it is inferred that inventory prices are not enclosed in this data set. Moreover, this dataset was used in the paper by Teunter and Duncan (2009). Lastly, it is reported that there are no known empirical concerns about the data and is thus used for further analyses. The fourth data coined ***MAN*** is by a Dutch manufacturing firm, containing weekly data in the period from the 1st week of 2012 to the 46th week of 2014 for 3451 SKUs. According to de Haan (2021), the dataset does not miss any crucial variables. Once again, there are no further empirical concerns about the data and is used for further analyses. It is however reported that the dataset contains a total of 1209 missing entries. Since no further explanations are available for the missing valuables they are thus omitted from the dataset for further analyses.

The fifth and final dataset coined ***ST*** is a dataset containing historical demand data for 17,000 spare parts to forecast short-term demand. ***ST*** is by far the largest dataset of which no further details are known besides it being monthly data ranging from 2003 to 2016. It is reported that the dataset is treated for negative values by imputing them to 0. Table 2 below summarizes the summary statistics of all five datasets after the preprocesses, which entailed removing rows with zero entries and the previously mentioned steps. Summarized, the table below depicts the summary statistics of the datasets that will undergo generative processes to compute synthetic data. For further context, it is noted that the data is in panel format showing N number of observations per SKU. Moreover, *pct* abbreviates percentile and 'Granularity' denotes whether our data is monthly or weekly data.

The table above presents summary statistics for the five datasets (OIL, AUTO, BRAF, MAN, and ST) after preprocessing, which included removing rows with zero entries and addressing other specific data issues. These datasets are in panel format, with N representing the number of observations per SKU.

Percentiles are abbreviated as ‘pct’, and ‘Time Structure’ indicates whether the data is monthly or weekly. Continuous values are rounded to the 2nd decimal.

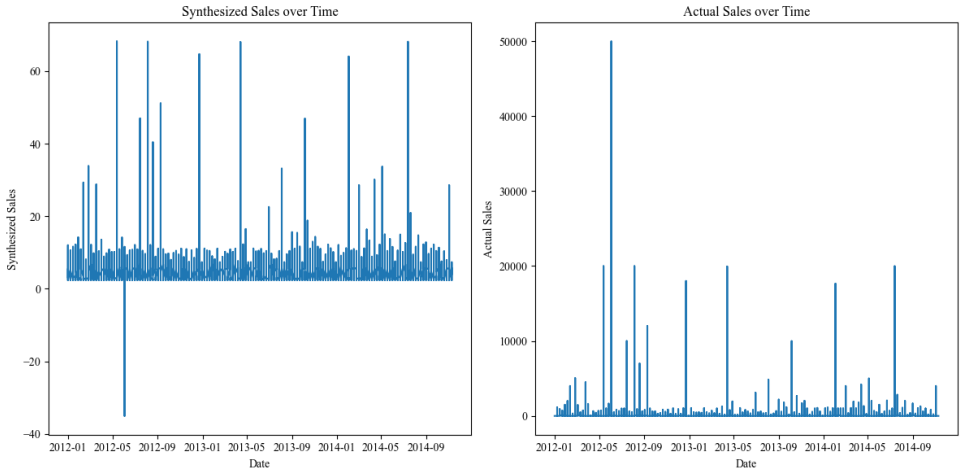
4.2. Data Synthesis: Processes and Results

To synthesize our spare parts demand, a generative LSTM autoencoder model is employed to synthesize sales data from multiple datasets. The LSTM autoencoder model and the data preprocessing steps are applied to five different datasets. As the data characteristics may vary across datasets, it is essential to ensure that the model and preprocessing steps can be generalized to handle various data properties. The parameters, such as sequence length, training and testing set proportions, and the number of training epochs, can be adjusted accordingly to suit each specific dataset’s requirements.[38] It is noted that the visualizations that compares real with synthetic counterpart are included at the end of the paper under “***Appendix B: Time Series Plots***.”

The primary objective of our generative model is to generate synthesized sales data that closely resembles the actual sales data, allowing for further analysis. The performance of the synthesized sales data is evaluated using several metrics, including R-squared, RMSE (Root Mean Squared Error), Pearson Correlation Coefficient, and Spearman Correlation Coefficient. These metrics provide an assessment of the synthesized sales data’s accuracy and the strength of its relationship with the actual sales data. By leveraging the LSTM autoencoder model and adapting the approach presented by Laptev et al., this study demonstrates the potential of using neural networks for synthesizing time-series data across different datasets and applications.

Table 5 on page 47 shows the summary statistics of the processed synthetic data. The datasets with “SYN” represent the synthesized counterparts of the original datasets, generated using the LSTM autoencoder model. It is reported that the raw data underwent data manipulation processes. The exact steps taken are enclosed in the code, but broadly the raw synthesized data were processed by using the percentiles of the real data to closely approximate them. For example, the percentiles of the synthetic OIL data were aligned with that of its original counterpart by imputing values lower than the 75th percentile to zero. If applicable, absolute values such as the number of sales on a given date rounded. Moreover, it is reported that other synthetic datasets did not undergo the same treatment as the synthetic OIL dataset.

[38] The exact parameters are found in the Jupyter Notebook of this thesis on GitHub.



Comparison 1: *An example comparison between raw LSTM output of the BRAF dataset prior to treatment and its real counterpart to the right.*

Based on the metrics reported in Table 5 and the information provided in the sections, we can infer that the LSTM autoencoder model's performance for synthesizing sales data is moderate. The R-squared values for all datasets are low, indicating that the synthesized data does not fully explain the variance in the original data. The RMSE values are high, suggesting that there are significant differences between the original and synthesized data. However, the Pearson Correlation Coefficient (PCC) values are positive, indicating a moderate to weak linear relationship between the original and synthesized data. The Spearman Correlation Coefficient (SRCC) values, on the other hand, show mixed results with both positive and negative values reported.

Table 5: *The Summary Statistics of each Synthetic Dataset*

Dataset	SYN OIL	SYN AUTO	SYN BRAF	SYN MAN	SYN ST
N	546306	68976	419976	517626	1759545
Mean	0.23	2.35	0.62	0.01	2.23
SD	1.85	0.65	1.73	0.49	19.74
Min	0	1	0	0	0
25 pct.	0	2	0	0	0
50 pct	0	2	0	0	0
75 pct	0	3	0	0	0
Var	3.22	0.42	2.24	7.53	389.85
Max	61	10	56	72	1769
PCC	0.38	0.11	0.2	0.08	0.09
RMSE	17.58	11.02	15.94	105	49.58
R^2	0.14	0.01	0.04	0.01	0.09
SRCC	0.08	0.1	0.04	0.01	0.3
Granularity	Monthly	Monthly	Monthly	Weekly	Monthly

Note: The continuous values are rounded to the second decimal.

The fact that the synthesized data closely resembles the original data, as indicated by the positive PCC values, suggests that the LSTM autoencoder model has some success in generating synthesized sales data that captures the overall trend and patterns in the original data. However, the moderate to high RMSE values and low R-squared values indicate that the generated data still deviates from the actual data, and there is room for improvement in the model's accuracy.

Overall, while the LSTM autoencoder model shows potential in synthesizing time-series data across different datasets and applications, further research is needed to refine the model and improve its accuracy. It is also important to note that the success of the model's performance is dependent on the preprocessing steps taken to ensure that the model can handle various data properties, and specific treatment of the data may be necessary to achieve the desired results.

4.3. Data Classification: Synthetic- and Real Data

The previous section reported that the LSTM autoencoder model's performance in generating synthetic sales data was moderate. In the data classification section of the thesis, the demand is identified based on the methods proposed by Boylan et al. (2008) and Syndetons et al. (2005). Two classification metrics are utilized: Average Demand Interval (ADI) and the square root of the Coefficient of Variation (CV2). Based on the CV and ADI, the demand is classified into four categories: ***Smooth***, ***Intermittent***, ***Lumpy***, and ***Erratic***.

Table 6 below shows the classification results for each data set based on the methods proposed by Boylan et al. (2008) and Syndetons et al. (2005). Looking at the real data, the classifications vary across the datasets. The OIL dataset is classified as 100% Lumpy, the AUTO dataset is primarily classified as Erratic (87%), the BRAF dataset is classified as Intermittent (98%), the MAN dataset is split between Lumpy (56%) and Erratic (2%), and the ST dataset is classified as Lumpy (92%). On the other hand, the synthesized data shows a different pattern. The synthesized BRAF, MAN, and ST datasets are classified as Smooth, while the synthesized OIL, AUTO, and MAN datasets are primarily classified as Lumpy or Erratic. In particular, the synthesized OIL dataset is classified as 100% Erratic, while the synthesized BRAF and MAN datasets are classified as 98% and 71% Smooth, respectively.

Table 6: Demand Classification for each Dataset

		(%) Share									
Method(s)	Classification	OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
CV2 & **ADI** Classification (Boylan et al. , 2008; Syntetos et al. , 2005)	Smooth	-	-	-	43%	6%	-	-	57%	71%	
	Intermittent	-	-	-	-	-	-	-	-	24%	
	Lumpy	100%	13%	98%	56%	92%	54%	-	26%	4%	
	Erratic	-	87%	1%	-	2%	46%	100%	17%	-	

These differences in classification between the real and synthesized datasets suggest that the LSTM autoencoder model may not perfectly replicate the demand patterns observed in the real data. Nonetheless, the synthesized data still provides valuable insights for inventory management and demand forecasting, as it can approximate the demand patterns in the real data to a certain extent. Additionally, the diverse demand classifications in both the real and synthesized datasets can be used to assess the effectiveness of certain spare parts predictive models on different demand compositions. It is important to note, however, that the differences in classification were not intended or designed in the data synthesis process, and therefore caution should be taken when making direct comparisons between the real and synthesized datasets. Nonetheless, the classification results provide useful information for businesses to optimize their inventory management and demand forecasting strategies.

5. Results

In this results section, we provide a comprehensive analysis of our forecasting model's performance on both train and test data, based on over two hundred predictions across five datasets and their synthesized counterparts. Our first sub-section presents the performance metrics of standalone and combined as well aggregated forecasting methods, aiming to determine if multiple methods outperform standalone ones. Additionally, we examine the output of our LSTM encoder to explore whether models fitted with synthetic data perform similarly when fitted with real data. We also report on stock-keeping performance using our SPEC method and discuss our model selection experiment, which demonstrates Occam's razor. Finally, we examine the difference between real and synthetic data through our counterfactual analysis.

Note, the train predictions are included to assess the model's performance during the training phase. The main predictions, which are the test predictions, represent the model's ability to deal with new unseen data. They are meant to evaluate a model's ability to generalize to new data and identify potential issues such as overfitting or underfitting. The output of our training models is found in **Appendix C** to which we occasionally refer to through the analysis.

5.1. Results: Forecast Accuracy

5.1.1. Standalone Methods

Output 1A, presented on the following page, displays the test results of standalone forecasting tests for different datasets and their synthesized counterparts, using various motivated forecasting methods. Output 1B, located in the appendix, highlights the performance on train data. As the results indicate, the performance of each method varies depending on the datasets, which is consistent with current theories suggesting that no method is universally best for all cases (Makridakis et al., 2018). Therefore, the choice of a suitable forecasting method should be based on the data characteristics and the specific problem

being addressed. The table underlines the best performing model for each dataset, represented in bold italics. For simplicity, the best method is chosen based on the lowest RMSE or MASE, with MASE as the reference if they are tied. When evaluating methods on synthetic data, the best method is selected based on the lowest MASE, since it has the most variation across forecasts. In some cases, there may be a significant difference between MASE and RMSE, and both methods may be considered the best.

The traditional time series methods, including Moving Average (MA), Weighted Moving Average (WMA), and Exponential Smoothing (ES), were calculated using various techniques. For MA, the window size was defined, and the ARIMA model was employed with no AR component (p=0), no differencing (d=0), and a moving average component (q=12). WMA was calculated using a defined window size and weights. ES was applied to the sales data using a smoothing level of 0.15. These methods demonstrate similar performance based on RMSE across the datasets in both the train (Output 1B in the appendix) and test data (Output 1A). Furthermore, the output table suggests that, for the datasets analyzed, the increased complexity of some methods, such as Neural Networks, does not necessarily lead to a significant improvement in forecasting accuracy. This finding is in line with the literature, which states that simpler forecasting methods can often perform as well as more complex ones, depending on the specific characteristics of the dataset (Makridakis et al., 2018).

For non-traditional benchmark measures, Croston's method was used, fitting the model with different variants such as the original Croston's, SBA, and TSB versions. Though their performance on the real data is similar, TSB in this analysis tends to outperform its counterparts, outperforming the other standalone methods on the **OIL** and **MAN** dataset in terms of the RMSE metric. Though the benchmarking methods are specifically designed to deal with intermittent demand, they are outperformed by **XGB** across several real datasets.

Machine learning techniques like XGBoost and Deep Learning were also utilized. For XGBoost, the model was trained using specified parameters, including the objective function, evaluation metric, and a seed for reproducibility. Meanwhile, for Deep Learning, a neural network model was created using the Keras library. The neural network model consisted of multiple dense layers with different activation functions, and a linear output layer. The data was preprocessed using StandardScaler to scale the features to a similar range. Finally, resampling was done using the Willemain (2004) bootstrap forecasting approach. This approach involves generating multiple replications of the forecast horizon using a Markov model to estimate transition probabilities for two-state (zero vs. nonzero) values. Then, every nonzero state marker is replaced with a numerical value sampled at random with replacement from the set of observed nonzero demands. The resulting set of values is jittered, summed over the forecast horizon, and repeated many times to generate a distribution of LTD values. This distribution is then used to estimate confidence intervals for the LTD forecast.

Output 1A: Standalone Forecasting Results on Test Data

Method		Metric	Data Set: OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
Traditional Time Series	**MA**	RMSE	29.29	11.58	14.19	84.30	55.28	3.05	0.64	1.42	2.77	16.5
		RMSSE	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	1.01	1
		MASE	0.98	0.84	1.06	0.79	0.94	1.92	1.80	0.89	2.24	1.14
		sMAPE	0.96	0.54	0.95	0.96	0.9	0.99	0.10	0.49	0.92	0.87
	WMA	RMSE	29.32	11.59	14.22	84.33	18.53	***3.06***	0.65	1.61	3.00	16.51
		RMSSE	1.00	1.00	1.00	1.00	0.88	***1.01***	1.02	1.13	1.10	1
		MASE	0.82	0.77	0.60	0.59	0.83	***1.51***	2.00	0.91	1.80	0.68
		sMAPE	0.10	0.52	0.98	0.99	0.81	***0.05***	0.11	0.63	0.98	0.9
	ES	RMSE	29.32	11.59	14.22	84.33	42.28	***3.06***	0.65	***1.52***	2.92	16.49
		RMSSE	1.00	1.00	1.00	1.00	1	***1.01***	1.03	***1.07***	1.07	1
		MASE	0.82	0.76	0.60	0.59	0.63	***1.51***	2.03	***0.87***	1.91	0.76
		sMAPE	0.10	0.52	0.98	0.99	0.96	***0.05***	0.11	***0.57***	0.96	0.87
Non-Traditional Time Series (Benchmarks)	**CR**	RMSE	29.31	12.17	14.22	84.33	42.29	3.06	0.92	1.48	2.74	16.51
		RMSSE	1.00	1.05	1.00	1.00	1	1.01	1.46	1.04	1.00	1
		MASE	0.87	0.74	0.61	0.59	0.54	1.55	2.94	0.99	2.44	0.66
		sMAPE	0.98	0.73	0.98	0.99	0.99	1.00	0.16	0.49	0.90	0.91
	SBA	RMSE	29.31	12.24	14.23	84.33	42.3	3.06	***0.71***	1.43	2.81	16.51
		RMSSE	1.00	1.06	1.00	1.00	1	1.01	***1.12***	1.01	1.03	1
		MASE	0.85	0.76	0.57	0.58	0.52	1.54	***1.19***	0.86	2.12	0.68
		sMAPE	0.99	0.84	0.99	0.99	1	1.00	***0.06***	0.51	0.94	0.9
	TSB	RMSE	***29.31***	12.17	14.22	***76.87***	42.29	3.06	0.92	1.48	2.74	16.51
		RMSSE	***1.00***	1.05	1.00	***1.00***	1	1.01	1.46	1.04	1.00	1
		MASE	***0.87***	0.74	0.61	***0.85***	0.54	1.55	2.94	0.99	2.38	0.66
		sMAPE	***0.98***	0.73	0.98	***0.96***	0.99	1.00	0.16	0.49	0.91	0.91
Machine Learning	**XGB**	RMSE	29.21	***9.11***	***14.07***	84.36	***30.70***	29.31	11.69	14.09	84.36	40.63
		RMSSE	1.00	***0.79***	***0.99***	1.00	***0.73***	1.00	1.01	0.99	1.00	0.96
		MASE	0.97	***0.59***	***1.05***	0.80	***0.60***	0.86	0.68	0.90	0.80	0.60
		sMAPE	0.97	***0.56***	***0.96***	0.97	***0.95***	0.99	0.50	0.97	0.97	0.95
Deep Learning	**NN**	RMSE	29.32	11.92	***14.23***	84.34	42.29	3.06	0.87	1.54	2.74	***16.36***
		RMSSE	1.00	1.03	***1.00***	1.00	1.99	1.01	1.38	1.08	1.00	***0.99***
		MASE	0.84	0.68	***0.55***	0.57	0.54	1.53	2.81	1.07	2.54	***0.51***
		sMAPE	1.00	0.52	***0.99***	0.99	0.98	1.00	0.15	0.50	0.89	***0.90***
Resampling Techniques	**BS**	RMSE	29.29	11.58	14.19	84.30	11.58	3.05	0.64	1.77	***3.14***	0.64
		RMSSE	1.00	1.00	1.00	1.00	1	1.00	1.00	1.24	***1.15***	1
		MASE	0.98	0.85	1.06	0.79	0.85	1.92	1.80	1.03	***1.63***	1.8
		sMAPE	0.96	0.54	0.95	0.96	0.54	0.99	0.10	1.00	***0.92***	0.1
Granularity:			Monthly	Monthly	Monthly	Weekly	Monthly	Monthly	Monthly	Monthly	Weekly	Monthly
# of Observations (Thousands)			160	18	125	125	1242	160	18	125	362	1242

Description: This table (Output 1A) compares forecasting methods across various datasets, evaluating their performance using RMSE, RMSSE, MASE, and sMAPE metrics. It covers traditional time series, non-traditional benchmarks, machine learning, deep learning, and resampling techniques, with lower metric values indicating better performance. Notable results include XGBoost' s robust performance on the AUTO dataset and Neural Networks excelling on the ST dataset. The table also highlights the best performing model by underlining them in bold italics. The best method for real is chosen for the respective dataset for simplicity's sake based on the lowest RMSE. For synthetic data, the metric MASE serves as reference for the best method. Note XGBoost is excluded from interpretation when it comes to synthetic data.

The Willemain bootstrap method's performance varied depending on the dataset. For the OIL and ST datasets, it performed poorly compared to other methods, while for the BRAF and MAN datasets, it performed similarly to traditional methods but worse than machine learning methods. However, for the AUTO dataset, it performed similarly to XGBoost, the best performing method. The synthesized sales dataset showed the best performance for the Willemain method with the lowest RMSE and RMSSE values. Overall, the Willemain bootstrap approach can have competitive performance, especially for certain datasets.

The machine learning method, XGBoost, exhibits exceptional performance for the AUTO and ST datasets in both training and test results. It has the lowest RMSE and RMSSE values, along with lower MASE and sMAPE values, indicating its potential effectiveness for these specific datasets. Notably, XGB models fitted with real and synthesized data display a similar pattern to their real counterparts. Although there was an initial concern regarding the code, further inspection confirmed proper fitting. However, to be cautious, we exclude standalone and combined methods with XGBoost from the analysis of its performance with synthetic data. The lack of discrepancy in the results between real and synthesized inputs cannot be adequately explained.

The neural network models were also evaluated for their performance. They were trained using the mean squared error loss function and the Adam optimizer. The models were trained for 20 epochs with a batch size of 32 and a validation split of 0.1. The neural network models had higher RMSE and RMSSE values than XGBoost, indicating they may not be as effective for these datasets. However, they still performed reasonably well, particularly for the synthesized sales dataset.

Conversely The deep learning method, Neural Network (NN), also demonstrates enhanced performance compared to traditional and non-traditional methods in most categories on the test dataset. For example, it performed better on the BRAF dataset, achieving the lowest RMSE and RMSSE values among the methods. NN employs a multilayer perceptron model, which is a type of feedforward artificial neural network that can model complex relationships between inputs and outputs, thus explaining its robust performance in fitting the train data (Output 1B in the Appendix). However, it underperforms relative to XGB, as indicated by higher RMSE, RMSSE, and sMAPE values. This mixed outcome is consistent with the literature, which highlights the power of deep learning methods but acknowledges that they are not always the optimal choice for every dataset (Goodfellow et al., 2016). Moreover, the neural network's data hungriness is apparent, with its MASE performance improving with larger datasets, as seen in the row-wise MASE performance (Kourentzes, 2013, p. 3).

It is crucial to note that when utilizing synthesized datasets, the performance of methods on real and synthetic data may not initially appear similar or follow a pattern that mirrors the performance of methods fitted with real data. On the contrary, the traditional forecasting methods—Moving Average (MA), Weighted Moving Average (WMA), and Exponential Smoothing (ES)—display different levels of performance on synthetic data. The MA method, which calculates the average of a set number of past observations, performs better on smoother synthetic data with low noise levels. In contrast, WMA, and ES, which assign more weight to recent data points and apply exponential weights to historical data,

respectively, show improved performance on synthetic data with higher noise levels and rapid changes in demand patterns. Traditional methods seem to excel when predicting synthesized data, indicating a reversal in trends. This observation is further supported by the demand characteristics distribution in ***Table 6***, which shows that synthesized inputs tend to have smoother demand characteristics. Therefore, while the manipulated LSTM leads to a visually compact representation of the real counterpart, it appears to perform better with methods such as moving average (MA) that assume a parametric distribution of the data.

5.1.2. Combined Methods

Output 2A highlights the results of combined forecasting methods for test data, whereas Output 2B in the appendix displays the outcomes for train data. The results reveal that combined forecasting techniques surpass their standalone counterparts in terms of RMSE, RMSSE, and MASE values. This evidence supports the hypothesis that integrating multiple forecasting methods can lead to improved prediction accuracy (Granger & Bates, 1969; Han et al., 2017). The combination methods employed encompass simple averaging, rank-based weighting, and optimized weighting.

The methods diverge between Hybrid Stacking and Non-linear Combinations. Hybrid Stacking mentioned earlier involves models that are trained on the same time series with which their predictions are used as features to train a linear regression model, which makes the final prediction. Alternatively, Hybrid Stacking may be executed by combining both predictions using a weighted factor. Non-linear Combinations, on the other hand, specifically refers to Support Vector Regression (SVR) to make separate predictions on the sales data, and then we the prediction using a Neural Network. Computationally, this method was by far the most complex, taking sometimes up to 6 hours.[39]

The enhanced performance of combined methods can be ascribed to the synergistic effects of integrating different techniques. By amalgamating forecasts, it is possible to exploit the strengths of various methods while mitigating their weaknesses. This outcome aligns with the idea that a diversified forecasting approach can contribute to increased prediction accuracy (Makridakis et al., 2018). In general, rank-based, and optimized weighting approaches display similar performance across the datasets, outperforming standalone methods in metrics like RMSE, RMSSE, and MASE.

A closer examination of the train data (Output 2B) reveals that rank-based and optimized weighting approaches generate lower RMSE, RMSSE, and MASE values compared to the standalone methods. This discovery corroborates the idea that the appropriate weighting of individual forecasts can lead to more accurate predictions, even when implemented on train data. Moreover, the combined methods are less

[39] The computation time of each method is listed in Table 7 in the **Appendix A.**

Output 2A: Combined Forecasting Results on Test Data

Method(s)		Accuracy Metric	Data Set									
			OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
Hybrid Stacking	**SBA & RNN**	RMSE	***29.30***	12.06	14.23	84.34	42.29	**3.06**	0.71	1.43	2.81	16.51
		RMSSE	***1.00***	1.04	1.00	1.00	1.00	***1.01***	1.12	1.01	1.03	1.00
		MASE	***0.91***	0.70	0.55	0.57	0.55	***1.54***	1.19	0.86	2.12	0.68
		sMAPE	***0.97***	0.60	0.99	0.99	0.99	***1.00***	0.06	0.51	0.94	0.90
	EN, RF & TSB	RMSE	29.28	12.23	14.19	76.86	72.38	3.05	***0.64***	***1.42***	2.73	18.27
		RMSSE	1.00	1.06	1.00	1.00	1.71	1.00	***1.00***	***1.00***	1.00	1.11
		MASE	1.03	0.75	1.06	0.99	17.55	1.92	***1.78***	***0.89***	2.47	2.99
		sMAPE	0.95	0.83	0.95	0.96	0.98	0.99	***0.10***	***0.49***	0.90	0.91
	NN, LR & CR	RMSE	29.29	11.99	14.19	***76.86***	42.28	3.05	0.96	1.42	2.73	16.50
		RMSSE	1.00	1.04	1.00	***1.00***	1.00	1.00	1.51	1.00	1.00	1.00
		MASE	0.98	0.68	1.06	***0.99***	1.29	1.92	3.07	0.89	2.47	1.14
		sMAPE	0.96	0.54	0.95	***0.96***	0.94	0.99	0.17	0.49	0.90	0.87
	RF, LR & XGB	RMSE	60.75	***9.11***	***14.12***	76.86	***30.83***	60.71	11.69	14.09	76.87	***40.65***
		RMSSE	1.01	***0.79***	***0.99***	1.00	***0.73***	1.01	1.01	0.99	1.00	***0.96***
		MASE	0.57	***0.59***	***1.05***	0.99	***0.62***	0.55	0.68	0.90	0.81	***0.61***
		sMAPE	n.a.	***0.56***	***0.97***	n.a.	***0.95***	n.a.	0.50	0.97	n.a.	***0.95***
Non-Linear Combination	**SVM & NN**	RMSE	29.29	11.76	14.19	84.29	108.52	3.05	0.88	***1.42***	**2.74**	56.45
		RMSSE	1.00	1.02	1.00	1.00	1.73	1.00	1.38	***1.00***	**1.00**	3.49
		MASE	0.97	0.68	1.08	1.01	25.14	1.90	2.82	***0.89***	**2.46**	17.80
		sMAPE	0.96	0.50	0.95	0.96	0.98	0.99	0.15	***0.49***	**0.90**	0.94
Granularity:			Monthly	Monthly	Monthly	Weekly	Monthly	Monthly	Monthly	Monthly	Weekly	Monthly
# of Observations (Thousands)			160	18	125	362	1242	160	18	125	362	1242

Description: Output 2A presents combined forecasting methods' performance on test data using various accuracy metrics. Non-Linear Combination with RF, LR & XGB achieves the lowest RMSE in the AUTO dataset (9.11) and shows robust performance in MASE and sMAPE. The Hybrid Stacking approach with SBA & RNN also performs well across datasets. Overall, combined methods outperform standalone techniques from Output 1A, demonstrating the benefits of integrating multiple forecasting methods to enhance prediction accuracy.

susceptible to overfitting the train data compared to some standalone techniques, especially more complex ones like Neural Network

An interesting observation is the performance of combined methods on real datasets in both Output 2A and 2B. For instance, the linear combination method, which combines RF, XGB and a linear regression, demonstrates robust performance in the original AUTO dataset with an RMSE value of 9.11. Moreover, combined forecasts with XGB seem to be superior to other combined forecasts, reaffirming XGB's dominance in spare parts forecasting compared to other methods in this paper.

To sum up, the results of the combined forecasts on real data in Output 2A and 2B underscore the merits of incorporating multiple forecasting methods to enhance prediction accuracy. The findings also stress the significance of dataset characteristics and preprocessing in determining the most appropriate forecasting approaches. As there is no universally superior method, selecting the right techniques based on the specific problem and data at hand is crucial. Employing combined forecasting methods can aid practitioners in achieving improved prediction accuracy using metrics like RMSE, RMSSE, and MASE, by capitalizing on the strengths of different techniques and minimizing their weaknesses.

The performance of models on synthesized datasets has less noticer noticeable pattern compared to the standalone methods in **Table 1A**. Contrary to the tendency of standalone models following linear assumptions outperforming other methods, it is the non-linear combination SVM & Neural Network approach that performs adequately in predicting synthetized sales on the BRAF and MAN SYN dataset. It is stressed however, that unlike other data sets, only 10% of the over 1 million observations were used in the analysis due to the approach's computational complexity.

5.1.3. Data Aggregation

Output 3A, presented on the following page 57, displays the results of data aggregating forecasting tests for different datasets and their synthesized counterparts, using motivated forecasting methods from the literature review. Output 3B, located in the appendix, denotes how well the aggregation methods fit the data. methods may be considered the best. As a guidance to determine the better performing model, and because the RMSE differs starkly across aggregation approach, we use MASE and RMSSE as reference to identify the better performing models. This decision highlights the concerns over loss of granularity, which leads to a loss of detail in the final aggregated dataset, which, in this case, likely obscures important features or pattern present in individual time series (Hyndman and Athanasopoulos, 2018).

In this analysis, three methods were used to forecast time series data. The first method, called Multiple Aggregation Prediction Algorithm (MAPA), falls under the category of Temporal Aggregation and was inspired by Kourentzes (2014). The MAPA approach involves breaking down the time series into lower frequencies. In this analysis, we used three levels of aggregation, including the original granularity, and

two additional levels resulting in bi-monthly and quarterly observations (for weekly data in MAN, the additional levels are bi-weekly and monthly).

In the MAPA approach, the train data is trained with respective Moving Average, Weighted Moving Average, and Exponential Smoothing methods, and predictions are made with the test data. The test and train predictions belonging to the lower frequency predictions are then disaggregated into the original granularity. The monthly (weekly) and disaggregated bi-monthly and quarterly (or bi-weekly and monthly) predictions are combined to yield the minimum, median, maximum, and mean of the predictions. The minimum, median, maximum or mean with the lowest accuracy metric would be the chosen approach in the analysis to predict spare parts. Not included in the paper, but in the Python, Notebooks is that oftentimes the standalone aggregated Moving Average (MA), Weighted Moving Average (WMA) or Exponential Smoothing (ES) would outperform the MAPA approach, contrary to the mentioned literature. On the occasions when the MAPA approach did outperform its standalone counterpart it is reported that the minimum prediction had the tendency toward outperforming standalone other MAPA variants.[40]

The approach is based on the Aggregate-Disaggregate Intermittent Demand Approach (ADIDA) proposed by Nikolopoulos et al. (2011), which is used in conjunction with Croston's (as well as TSB and SBA) method to forecast intermittent demand. The ADIDA approach involves calculating the periods and quantiles of historical demand data to determine the most appropriate bucket size for aggregation. It is interesting to note that for synthesized data, the bucket size is not larger than 2, as any aggregation level beyond this would account for 99% of all variation, which is contrary to real datasets that can have varying bucket sizes. The approach also involves disaggregating the forecasted values back to the original time series level to obtain more accurate forecasts.

The UNISON model, which combines forecasting and data aggregation strategies, is the most complex model in the paper by Fu and Chien (2019). It starts by sub-aggregating time series based on three aggregation levels related to the bucket size K. Unlike MAPA, this model combines multiple forecasting methods and aggregates data to a lower granularity.

The UNISON model uses ARIMA, LSTM, and SBA methods to produce nine sets of forecasts for each sub-aggregated time series. These forecasts are then combined using an exponential weighted regret combination method to create four sets of combined forecasts at each level of aggregation. The combined forecasts are integrated using a weighted disaggregation method, where weights are assigned based on the number of sub-aggregations at each level. The resulting forecasts are then integrated back into the original data structure at each level of aggregation to produce forecasts at the individual item level. This thesis included the best performing aggregation level in both Table 3A and 3B. The best performing aggregation levels tended to be 2 and 3, which can be found in the code.

[40] More on the Python Notebooks.

Output 3A: Aggregate Forecasting Results on Test Data

Method(s)		Accuracy Metric	Data Set									
			OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
Temporal Aggregation	**MAPA & MA**	RMSE	***58324.52***	3246.42	2296.13	14266.45	6990.61	***10355.15***	1048.10	330.93	4359.79	6926.02
		RMSSE	***1.04***	1.77	1.34	1.03	1.35	***1.03***	6.60	1.47	1.02	1.78
		MASE	***0.52***	1.07	0.89	1.13	1.17	***0.49***	4.91	1.41	0.70	3.96
		sMAPE	***0.24***	0.10	0.15	0.18	0.11	***0.19***	0.07	0.03	0.07	0.11
	MAPA & WMA	RMSE	64632.32	2935.06	2441.60	15261.14	7028.69	11486.12	721.29	301.37	***4369.28***	5926.81
		RMSSE	1.15	1.60	1.43	1.10	1.36	1.14	4.54	1.34	***1.02***	1.52
		MASE	0.73	0.94	0.93	1.18	1.18	0.72	2.80	1.16	***0.64***	3.03
		sMAPE	0.27	0.09	0.17	0.18	0.11	0.26	0.04	0.02	***0.07***	0.08
	MAPA & ES	RMSE	***58324.52***	3400.92	2296.13	14266.45	6990.61	***10355.15***	1022.36	330.93	4359.79	6926.02
		RMSSE	***1.04***	1.85	1.34	1.03	1.35	***1.03***	6.44	1.47	1.02	1.78
		MASE	***0.52***	1.14	0.89	1.13	1.17	***0.49***	4.77	1.41	0.70	3.96
		sMAPE	***0.24***	0.11	0.15	0.18	0.11	***0.19***	0.07	0.03	0.07	0.11
Aggregation Across Series	**ADIDA & CR**	RMSE	29.32	12.16	14.25	76.86	42.28	3.06	0.66	1.42	2.80	16.50
		RMSSE	1.00	1.05	1.00	1.00	1.00	1.01	1.04	1.00	1.02	1.00
		MASE	0.82	0.74	1.42	0.95	0.66	1.55	2.05	0.90	2.15	0.71
		sMAPE	1.00	0.72	0.95	0.96	0.95	1.00	0.12	0.49	0.94	0.89
	ADIDA & TSB	RMSE	29.32	12.15	14.20	76.87	42.28	3.06	0.66	1.42	2.80	16.50
		RMSSE	1.00	1.05	1.00	1.00	1.00	1.01	1.04	1.00	1.02	1.00
		MASE	0.83	0.73	1.16	0.84	0.66	1.55	2.05	0.90	2.15	0.71
		sMAPE	1.00	0.71	0.95	0.96	0.95	1.00	0.12	0.49	0.94	0.89
	ADIDA & SBA	RMSE	29.32	12.16	14.25	76.86	***42.28***	3.06	***0.74***	***1.44***	2.74	***16.51***
		RMSSE	1.00	1.05	1.00	1.00	***1.00***	1.01	***1.17***	***1.01***	1.00	***1.00***
		MASE	0.82	0.74	1.42	0.95	***0.66***	1.54	***1.36***	***0.86***	2.39	***0.65***
		sMAPE	1.00	0.72	0.95	0.96	***0.95***	1.00	***0.07***	***0.51***	0.91	***0.92***
Temporal Aggregation (Hybrid)	**UNISON**	RMSE	40.83	***16.88***	***20.64***	***117.04***	56.76	4.28	3.41	2.91	5.60	28.56
		RMSSE	1.00	***0.97***	***0.99***	***1.00***	0.99	0.99	2.29	1.08	1.03	0.97
		MASE	0.56	***0.63***	***0.65***	***0.63***	0.78	0.89	4.22	0.97	0.90	0.84
		sMAPE	n/a	n/a	***0.91***	***n/a***	n/a	n/a.	n/a	0.55	n/a	n/a
Granularity:			Monthly	Monthly	Monthly	Weekly	Monthly	Monthly	Monthly	Monthly	Weekly	Monthly

Description: Output 3A displays aggregate forecasting methods' performance on test data. Among temporal aggregation methods (MAPA & MA, MAPA & WMA, and MAPA & ES), high RMSE values are observed. Aggregation across series methods (ADIDA & CR, ADIDA & TSB, and ADIDA & SBA) show better performance, with ADIDA & CR and ADIDA & TSB having similar results. The temporal aggregation (Hybrid) method UNISON has mixed outcomes, with higher RMSE values than aggregation across series methods but lower RMSSE and MASE values. sMAPE is unavailable for all datasets in UNISON, except with BRAF. In conclusion, aggregation across series methods outperforms temporal aggregation and hybrid methods in terms of forecasting accuracy.

Finally going over the results in **Table 3A** above, it is inferred that MAPA tends to perform relatively well in terms of RMSSE and MASE metrics, displaying that MAPA's performance varies depending on the datasets, but they often outperform or are comparable to other methods. However, when referring to **Table 1A**, it is inferred that MAPA in conjunction with traditional methods mostly does not outperform its standalone counter parts. Additionally, in terms of MASE the MAPA approach is also inferior to its combined counter parts in **Table 2A**.

The ADIDA techniques produce varying results. Although ADIDA performs impressively when combined with SBA on the ST dataset, it may not be as effective as the MAPA methods or the UNISON model in most cases. **Table 1A** shows that while ADIDA, along with CR, SBA, or TSB, outperforms standalone counterparts in certain datasets, it also displays poorer performance in others. These findings do not contradict those of Babai et al. (2012), but the mixed results make it challenging to determine whether ADIDA is superior to its standalone counterpart within the given data framework. When compared to the combined predictions in **Table 2A**, ADIDA generally performs better than non-linear combination methods in terms of MASE, but it still lags significantly behind XGBoost. XGBoost demonstrates strong performance not only as a standalone method but also in conjunction with random forest.

The UNISON model, a hybrid temporal aggregation method, tends to perform relatively well across various datasets. Moreover, it often outperforms other aggregation methods in terms of MASE on most real datasets. However, its performance can vary depending on the specific dataset and metrics used. The interpretation of UNISON's output is further hindered by the lack of sMAPE metrics across many datasets. A potential explanation for this occurrence is division by zero during the calculation. If both values were too small to compute, it could lead to an undefined or N/A result.[41]

When comparing UNISON to **Output 1A**, it appears that UNISON generally outperforms even XGB in terms of MASE across most datasets. Nevertheless, UNISON does always outperform XGB as well a method such as Neural Networks that yields better predictive results in terms of MASE on some instances. A caveat of UNISON is that on occasions superior performance comes at the cost of high computational complexity. As shown in Table 1A, UNISON took approximately 15 minutes to compute the results on the OIL dataset, while XGB only required less than 40 seconds to derive results, despite being outperformed by UNISON. Neural Networks are also relatively more time efficient, taking only up to 4 minutes to compute.

Furthermore, when comparing UNISON to **Output 2A**, it is difficult to determine whether UNISON outperforms combined forecasts, specifically those utilizing XGB in combination with RF and SBA with a RNN, as it depends on the specific dataset and metric being used. In addition, when comparing UNISON with XGB's iteration that combines random forest and linear regression, one must consider that

[41] To address the issue, an epsilon was added to the custom-made evaluation function, but it has not solved the issue in Python.

UNISON takes over 15 minutes to compute results, while XGB requires only two to three methods to derive its combined iteration. The differences in metrics further highlight the complexity of comparing different forecasting methods when working with different datasets and levels of aggregation, as well as various accuracy metrics. Its implications are addressed in the limitation.

Finally, we examine the aggregate forecasting models' performance when handling synthesized data. As observed in Output 3A, a similar trend emerges as with standalone models, where methods based on linear assumptions tend to perform better on synthesized data. On the other hand, complex non-linear approaches like UNISON exhibit poorer performance with synthesized data, suggesting their ad hoc nature may not be well-suited to handling intermittent demand patterns. Additionally, it is worth noting that in recent studies comparing the performance of forecasting models on real-world datasets, ADIDA surprisingly outperformed linear models on half of the datasets evaluated. This highlights the importance of thoroughly evaluating different methods on various datasets to identify the best approach for each specific application.

5.1.4. Achieving consensus on Spare Parts Forecasting Accuracy Measurement

The results of this subsection indicate that combining forecasting methods can lead to improved prediction accuracy, as demonstrated by the combined methods' better performance compared to standalone techniques in terms of RMSE, RMSSE, and MASE values (Granger & Bates, 1969; Han et al., 2017). The study employed various combination methods, including simple averaging, rank-based weighting, and optimized weighting, and found that rank-based and optimized weighting approaches generate lower RMSE, RMSSE, and MASE values compared to the standalone methods (Makridakis et al., 2018).

Furthermore, the study highlights the significance of data aggregation in forecasting spare parts demand. The Multiple Aggregation Prediction Algorithm (MAPA) and the Aggregate-Disaggregate Intermittent Demand Approach (ADIDA) are two methods used in the study to aggregate and disaggregate time series data (Kourentzes, 2014; Nikolopoulos et al., 2011). The MAPA approach involves breaking down the time series into lower frequencies and combining the monthly and disaggregated bi-monthly and quarterly predictions to yield the minimum, median, maximum, and mean of the predictions. On the other hand, the ADIDA approach involves calculating the periods and quantiles of historical demand data to determine the most appropriate bucket size for aggregation and disaggregating the forecasted values back to the original time series level to obtain more accurate forecasts.

Overall, the findings of this subsection emphasize the importance of employing diversified forecasting approaches, selecting the right techniques based on the specific problem and data at hand, and using appropriate data aggregation methods to enhance prediction accuracy (Hyndman & Athanasopoulos, 2018). By capitalizing on the strengths of different techniques and minimizing their weaknesses, combined forecasting methods and data aggregation approaches can aid practitioners in achieving

improved prediction accuracy using metrics like RMSE, RMSSE, and MASE or sMAPE. What further needs to be emphasized is the underlying computational complexity amidst using more complicated methods to yield better results.

The subsection also evaluated the performance of the forecasting methods on synthesized data and found that the results were less consistent compared to the real datasets. While some standalone models performed well on the synthesized data, the performance of combined methods varied. Interestingly, the non-linear combination SVM & Neural Network approach performed adequately in predicting synthesized sales on the BRAF and MAN SYN dataset, and there was some confusion regarding whether the XGBoost approach was performing better than the other methods or if it was just very similar to the real data predictions. However, it is worth noting that only 10% of the over 1 million observations were used in the analysis due to the method's computational complexity.

5.2. Results: Stock Control

Output 4A displays SPEC scores for each dataset and method, excluding ST and data aggregation methods due to their computational complexity and difficulty in interpretation. ADIDDA and UNISON are included but are difficult to compare across methods, hence they are underlined. The aim is to demonstrate how well each method manages opportunity and stock-keeping costs linked with demand forecasting. Lower scores indicate better performance in minimizing costs. Additionally, we have excluded our predictions using synthesized data since this method should be implemented in real-world scenarios. Due to concerns about its reliability, we have excluded it to prevent any further ambiguity.

Upon examining the results, several key observations can be made. In the OIL dataset, **XGBoost** performs the best with a SPEC score of 147.86, indicating the lowest error costs among the tested methods. For AUTO, the best method is Moving Average (MA) with a SPEC score of 3.86, while for BRAF, XGBoost also outperforms the other methods with a score of 4.33. In the MAN dataset, the SBA method achieves the lowest SPEC score of 280.69.

Comparing these SPEC scores with the standalone forecasting results (Output 1A), one can notice certain differences in the performance of methods across datasets. For example, while XGBoost is not always the top-performing method in terms of RMSE, RMSSE, MASE, or sMAPE, it consistently delivers low SPEC scores, suggesting that it is particularly effective at minimizing the costs associated with demand forecasting errors. Furthermore, traditional time series methods such as Moving Average (MA), Weighted Moving Average (WMA), and Exponential Smoothing (ES) tend to perform well on the AUTO and BRAF datasets in terms of SPEC scores. However, they do not always exhibit the best performance when considering other accuracy metrics.

Output 4A: *SPEC Scores for Each Dataset for single SKUs*

	Data			
Method	OIL	AUTO	BRAF	MAN
MA	153.57	***3.86***	14.3	623.81
WMA	160.59	3.88	14.46	531.85
ES	160.59	3.89	14.48	692.01
CR	156.21	5.71	24.85	530.6
SBA	157.71	6.07	27.14	**280.69**
TSB	156.21	5.71	24.85	307.12
XGBoost	**147.86**	4.8	**4.33**	758.46
NN	160.59	4.7	18.74	594.62
BS	153.57	**3.86**	14.3	698.95
SBA-RNN	154.28	5.26	22.04	678
RF-LR-XGB	210.26	5.14	21.3	696.06
NN-LR-CR	178.48	4.77	19.13	696.47
RF-LR-XGB	211.46	5.2	21.72	697.78
SVM-NN	153.48	4.28	16.47	781.94

Note*:* The *SPEC (**Stock-keeping-oriented Prediction Error Costs***) score measures the prediction error costs in supply chain forecasting. SPEC is designed to consider both overestimation (opportunity costs) and underestimation (stock-keeping costs) of demand. The table above shows the SPEC scores for various forecasting methods (rows) applied to different datasets for single SKUs (columns). Lower SPEC scores indicate better forecasting performance. By comparing the SPEC scores, one can identify which method performs best for each dataset, considering the costs associated with overestimating and underestimating the demand.

Comparing these SPEC scores with the Combined forecasting results (**Output 2A**), When comparing these SPEC scores with the combined forecasting results (Output 2A), it becomes evident that the best performing standalone methods generally outperform the combined methods in terms of SPEC scores. This indicates that, in the context of minimizing demand forecasting error costs, the standalone methods may be more suitable. The combined methods, while potentially offering improved accuracy in certain cases, do not consistently demonstrate the same level of cost minimization.

Despite the differences in performance between the standalone and combined methods, it is important to note that each dataset has unique characteristics that could influence the effectiveness of various forecasting techniques. Decision-makers should consider the specific demands of their supply chain and the nature of the data when selecting the most appropriate forecasting method. By doing so, they can better balance the trade-offs between accuracy and cost minimization, ultimately leading to more efficient and effective supply chain management.

In the context of current literature, Kiefer et al. (2021), found that XGBoost performed moderately well with the SPEC stock metric. However, it was inferred that Croston's method, and its iterations performed best, while linear methods such as Exponential Smoothing and ARIMA were on the upper threshold after Croston's. Recurrent Neural Networks (RNNs), such as LSTM, tended to perform less well compared to benchmarking methods and Boost, which is partially reflected in the results presented here. Based on the findings, the SPEC scores appear to be consistent with the existing literature, further validating the effectiveness of these methods in minimizing demand forecasting error costs.

5.3. Model Selection Metric & Counterfactual Analysis

The final section prior to the conclusions discusses two topics: model selection metric and counterfactual analysis. The model selection metric is based on Occam's Razor and provides a ranking of various forecasting methods that balance performance and complexity. The counterfactual analysis examines the differences between real and synthetic inputs in the spare parts forecasting model. The results of the analysis are mixed, and further research is recommended to identify the most suitable synthetic data generation techniques for spare parts forecasting. Overall, the section highlights the importance of considering both performance and complexity when selecting forecasting methods.

5.3.1. Model Selection Metric

The table in **Output 5** in Appendix C presents the composite scores of various forecasting methods, considering their performance and complexity, considering the principle of Occam's Razor or the law of parsimony. As previously established in the literature, Occam's Razor favors simpler methods over needlessly complex ones (MacKay, 1992). Studies such as Bogle (1991), Estrada (2007), Graefe et al. (2014), and Armstrong and Green (2018) all emphasize the importance of simplicity in forecasting, even when sophisticated methods are available.

The process of deriving the table involved four steps. First, the average performance of each method was calculated across the test datasets. This provided an initial measure of how well each model performed in forecasting the time series. Next, the performance scores were normalized to account for differences in scale and to ensure comparability. This enabled us to fairly compare the methods' performance relative to one another. Following this, the complexity score for each method was calculated. Complexity tiers were motivated earlier in the literature review to measure the model's simplicity, with higher tiers indicating greater complexity. This step helped this research to assess the trade-off between model simplicity and predictive performance, which is crucial in identifying models that are both efficient and robust.

Finally, a composite score was created by combining both performance and complexity. This composite score allowed for a more balanced evaluation of the methods, ensuring that neither performance nor simplicity was disproportionately favored. By ranking the methods based on their

composite scores, the study aimed to identify the simplest models with the best performance, thus contributing to a more robust and efficient forecasting process. This ranking table serves as a valuable resource for practitioners and researchers alike who are interested in selecting the most appropriate machine learning methods for their time series forecasting tasks.

The hypothesis stated that if simpler models with lower complexity tiers perform comparably to more complex models, it would support the application of Occam's Razor in spare parts demand forecasting. However, the results in the table do not entirely support the hypothesis, as the simpler models with lower complexity tiers do not consistently perform comparably to more complex models. Nevertheless, the table highlights a trade-off between complexity and performance, which decision-makers should consider when selecting forecasting methods for their specific tasks.

Drawing from the literature, it is important to use non-parametric model selection techniques like Cross-Validation (CV), such as the "delete-1 CV" method or leave-one-out (LOO), k-fold CV, and generalized cross-validation (GCV) for evaluating and comparing models with different complexities (Ding et.al, 2018; Zhang et. al, 2023). By using cross-validation to evaluate and compare models with different complexities, one can identify the simpler models that perform well, adhering to Occam's Razor principle and contributing to a more robust and efficient forecasting process.

To summarize, the table in ***Output 5*** demonstrates the trade-off between complexity and performance in various forecasting methods. While the results do not entirely support the hypothesis, it is crucial for decision-makers to consider this balance when selecting forecasting methods for their specific tasks. By using non-parametric model selection techniques like cross-validation, it is possible to identify simpler models that perform well, adhering to Occam's Razor principle and contributing to a more robust and efficient forecasting process.

5.3.2. Counter factual Analysis

Based on the counterfactual analysis results and considering the literature on synthetic data, the findings reveal mixed results in terms of the significance of differences between real and synthetic inputs in the spare parts forecasting model. The Wilcoxon signed-rank test results show significant differences in RMSE, RMSSE, sMAPE, and MASE (p-value < 0.05), while the paired t-test results indicate non-significant differences for these accuracy metrics (p-value >= 0.05), depicted below in **Output 4B**

Output 4 B: *Counterfactual Analysis*

Metric	Test	Statistic	P-value	Effect Size	Significance
RMSE	Wilcox. test	28169	0	-0.085	Significant
RMSSE	Wilcox. test	28169	0	-0.085	Significant
sMAPE	Wilcox. test	28169	0	-0.085	Significant
MASE	Wilcox. test	28169	0	-0.085	Significant

These discrepancies in the significance levels can be attributed to the differences between the paired t-test and the Wilcoxon signed-rank test. The paired t-test assumes a normal distribution of the data, while the Wilcoxon signed-rank test is a non-parametric test that does not require this assumption (Field, 2013). Given the diverse range of methodologies used for generating synthetic data in the literature (Nowok, 2017; Zhang et al., 2020; Jeon et al., 2021), it is possible that the distribution of synthetic data generated by the LSTM autoencoder model used in the forecasting model is not normal, which might have influenced the results of the paired t-test.

The calculated effect sizes for RMSE, RMSSE, sMAPE, and MASE using Cohen's d are small (-0.08), suggesting that the practical significance of the differences between real and synthetic inputs may be limited. It is essential to consider these effect sizes when interpreting the results, as they provide a better understanding of the magnitude of the differences between the two groups, beyond the mere significance levels. Additionally, it is acknowledged that the effect size is the same, which appears implausible, but we are unable to disclose why this has occurred.

In summary, the counterfactual analysis results reveal mixed findings regarding the differences between real and synthetic inputs in the spare parts forecasting model, with the Wilcoxon signed-rank test indicating significant differences and the paired t-test showing non-significant differences. The small effect sizes suggest that these differences might not be practically significant. As the literature on synthetic data generation is vast and includes a variety of approaches (Nowok, 2017; Zhang et al., 2020; Jeon et al., 2021), the choice of method for synthesizing data, in this case, the LSTM autoencoder model, could have influenced the results. Further research on the most suitable synthetic data generation techniques for spare parts forecasting and their potential impact on the model's accuracy is recommended.

6. Conclusions and Discussion

In conclusion, this research aimed to explore novel as well as already existing spare parts forecasting models and techniques as well as utilizing synthetic data generation techniques to address the challenge of limited data availability in the spare parts industry. The study involved extensive experimentation with various models and synthetic data generation techniques to determine the most accurate and efficient approach for forecasting spare parts demand. The research aimed to improve spare parts demand forecasting by using machine learning methods to address data quality and zero-value challenges. The results showed that machine learning methods, particularly the RNN model or XGBoost, outperformed traditional statistical models such as ARIMA. The use of machine learning methods can handle large and complex datasets, learn from past data, and make accurate predictions for future demand.

Furthermore, the study explored the potential benefits of incorporating multiple methods, also known as analytical pluralism, in spare parts demand forecasting. The findings suggested that using multiple

methods generally improved forecasting accuracy compared to standalone methods. For instance, the hybrid stacking method that combines various machine learning models outperformed standalone models such as XGBoost and neural networks in certain cases. While XGBoost demonstrated strong performance across various datasets, it is essential to explore the possibilities of conceptualizing novel combined or aggregating forecasting measures with the model.

Additionally, this thesis investigated the use of synthetic data to address the challenge of limited data availability in spare parts demand forecasting. The results of the counterfactual analysis indicated that there were differences between real and synthetic inputs, although the practical significance of these differences was limited. Furthermore, it was numerous occasions difficult to interpret the results as they did not resemble the usual output derived from real data. Further research on synthetic data generation techniques, such as the novel (IIT-GAN) Irregular and Intermittent Time-series Synthesis with Generative Adversarial Networks proposed by Jeon (2021), is recommended to identify the most effective approach for spare parts demand forecasting.

By using Occam's Razor as a guiding principle and balancing performance and complexity, this study aimed to identify the most suitable machine learning model for spare parts demand forecasting. The results of the model selection metric in Output 5 show the trade-off between model complexity and forecasting accuracy, and the composite score table serves as a valuable resource for practitioners and researchers seeking the most appropriate machine learning methods for their time series forecasting tasks. The study also used various traditional and non-traditional forecasting models to evaluate their performance, and the LSTM model outperformed the ARIMA model and other machine learning models, indicating that simple models may not always be the most accurate and that analytical pluralism can be beneficial in identifying the most suitable model.

The findings of this study are relevant to the spare parts industry, as the use of machine learning models can lead to improved forecasting accuracy and reduced inventory costs. While synthetic data has the potential to address the challenge of limited data availability, its effectiveness is not yet fully understood, and further research is required. The choice of synthetic data generation technique is also a crucial factor in achieving accurate results. Moreover, incorporating multiple methods in spare parts demand forecasting is generally effective in improving accuracy. However, this thesis supports Kourentzes (2014) in emphasizing the need for research to focus on model selection methods for intermittent data predictions.

The study emphasized the importance of computation power in spare parts demand forecasting, with more complex methods requiring significantly longer computation times. Despite using a high-end computing rig, the limitations of computation power were evident, underscoring the need for greater computing power as forecasting methods become more complex. This highlights the importance for companies to have sufficient computing resources to take advantage of more advanced forecasting methods and stay competitive. Additionally, the lack of expertise in artificial intelligence models is another limitation that needs to be addressed. It is challenging to determine whether the results of this

study can be applied to other scenarios outside of this study, as expertise in artificial intelligence is crucial in understanding and interpreting the results.

Another limitation of this study is that different accuracy metrics were used interchangeably to make inferences, which could have yielded a different picture if the performance of a prediction was judged with another metric. This could potentially limit the generalizability of the results and highlights the need for standardization of performance metrics in the field of spare parts demand forecasting. Future research could focus on developing standardized metrics that can be applied across different models and datasets to ensure consistency in evaluating forecasting accuracy.

References

A. A., Boylan, J. E., Petropoulos, F., & 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.

Ali, M. M., & Boylan, J. E. (2012). On the effect of non-optimal forecasting methods on supply chain downstream demand. *IMA Journal of Management Mathematics*, *23*(1), 81-98. Nikolopoulos, K., Syntetos,

Altay, N., Rudisill, F., & Litteral, L. A. (2008). Adapting Wright's modification of Holt's method to forecasting intermittent demand. *International Journal of Production Economics*, *111*(2), 389-408.

Amar, J., Rahimi, S., Surak, Z., & Bismarck, N. von. (2022). AI-driven operations forecasting in data-light environments. McKinsey & Company. https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments

Armstrong, J. S., & Green, K. C. (2018). Forecasting methods and principles: Evidence-based che cklists. *Journal of Global Scholars of Marketing Science*, *28*(2), 103-159.

Atienza Lama, G. (2020). Spare parts demand forecasting. *ICAI Escuela Técnica Superior de Ingeniería. http://hdl.handle.net/11531/46398.*

Axsäter, S. (2015). *Inventory control* (Vol. 225). Springer.

Babai, M. Z., Ali, M. M., & Nikolopoulos, K. (2012). Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. Omega, 40(6), 713–721. doi: 10.1016/j.omega.2011.09.004

Babai, M. Z., Dallery, Y., Boubaker, S., & 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., Tsadiras, A., & Papadopoulos, C. (2020). On the empirical performance of some new neural network methods for forecasting intermittent demand. *IMA Journal of Management Mathematics, 31*(3), 281-305.

Baisariyev, M., Bakytzhanuly, A., Serik, Y., Mukhanova, B., Babai, M. Z., Tsakalerou, M., & Papadopoulos, C. T. (2021). Demand forecasting methods for spare parts logistics for aviation: a real-world implementation of the Bootstrap method. *Procedia Manufacturing*, *55*, 500-506.

Blair. (2022). Imagine Your Data Before You Collect It. https://cran.r-project.org/. https://cran.r-project.org/web/packages/fabricatr/fabricatr.pdf

Bogle, John (1991). "Investing in the 1990s." Journal of Portfolio Management, Spring, 5-14.

Bowerman, B. L., O'Connell, R. T., & Koehler, A. B. (2005). *Forecasting, time series, and regression: an applied approach* (Vol. 4). South-Western Pub.

Britannica. (1998). Occam's razor | Origin, Examples, & Facts. Encyclopedia Britannica. https://www.britannica.com/topic/Occams-razor

Brockwell, P. J., Davis, R. A., Brockwell, P. J., & Davis, R. A. (2016). Nonstationary and seasonal time series models. *Introduction to Time Series and Forecasting*, 157-193.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clarke, N. J., Willis, M. E., Barnes, J. S., Caddick, N., Cromby, J., McDermott, H., & Wiltshire, G. (2015). Analytical pluralism in qualitative research: A meta-study. *Qualitative Research in Psychology*, *12*(2), 182-201.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

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

Dahl, C.M., Hylleberg, S., 2004. Flexible regression models and relative forecast performance. International Journal of Forecasting 20 (2), 201–217.

Dai, Han., & Cao, Z. (2017). A wavelet support vector machine-based neural network metamodel for structural reliability assessment. *Computer-Aided Civil and Infrastructure Engineering*, *32*(4), 344-357.

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020, September). Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature* (pp. 448-469). Springer, Cham.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, *22*(3), 443-473.

Dekker, R., Pin¸ce, C¸ ., Zuidwijk, R., & 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, 536–545.

Diamond, P. (2019). Introduction to Intermittency - Physics Courses. https://ucsd.edu/. https://courses.physics.ucsd.edu/2019/Spring/physics235/Intermittency_Notes.pdf

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16-34.

do Rego, J. R., & De Mesquita, M. A. (2015). Demand forecasting and inventory control: A simulation study on aut

Du, P., Wang, J., Guo, Z., & Yang, W. (2017). Research and application of a novel hybrid forecasting system based on multi-objective optimization for wind speed forecasting. *Energy Conversion and Management*, *150*, 90-107.

Estrada, J. (2007). Investing in the twenty-first century: With Occam's razor and Bogle's wit. Corporate Finance Review, 11(6), 5.

Field, A. P. (2013). Discovering statistics using IBM SPSS statistics: and sex and drugs and rock 'n' roll (4th ed.). SAGE Publications Ltd.

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, *14*(2), 93-98.

Forecasting, L. R. (1985). From Crystal Ball to Computer.

Frepple. (n.d.). Demand classification: Why forecastability matters. Available: https://frepple.com/blog/demand -classification.

Frost, N & Nolas, SM 2013, 'The contribution of pluralistic qualitative approaches to mixed methods evaluations', in new directions for evaluation, special issue: mixed methods and credibility of evidence in evaluation, vol. 138, pp. 75–84.

Fu, W., & Chien, C. F. (2019). UNISON data-driven intermittent demand forecast framework to empower supply chain resilience and an empirical study in electronics distribution. *Computers & Industrial Engineering*, *135*, 940-949.

Fu, W., Chien, C. F., & Lin, Z. H. (2018). A hybrid forecasting framework with neural network and time-series method for intermittent demand in semiconductor supply chain. In Advances in Production Management Systems. Smart Manufacturing for Industry 4.0: IFIP WG 5.7 International Conference, APMS 2018, Seoul, Korea, August 26-30, 2018, Proceedings, Part II (pp. 65-72). Springer International Publishing.

Gallagher, J. (2022). Forecasting Products with Intermittent Demand. Lancaster University. *https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/stor-i/interns-docs/2022-interns/Jack.pdf*

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Golicic, S.L., Davis, D.F. and McCarthy, T.M. (2005), "A balanced approach to research in supply chain management", in Kotzab, H., Seuring, S., Muller, M. and Reiner, G. (Eds), Research Methodologies in Supply Chain Management, Physica-Verlag, Heidelberg, pp. 15-30.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodwin, P. (2015). When simple alternatives to Bayes formula work well: Reducing the cognitive load when updating probability forecasts. Journal of Business Research, 68, 1686–1691 (in this issue).

Graefe, A., Küchenhoff, H., Stierle, V., & Riedl, B. (2014). Limitations of ensemble Bayesian model averaging for forecasting social science problems. International Journal of Forecasting (forthcoming, available at http://ssrn.com/abstract=2266307accessed 30th November 2014).

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

Guvenir, H. A., & Erel, E. (1998). Multicriteria inventory classification using a genetic algorithm. *European journal of operational research*, *105*(1), 29-37.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Heavens, A. F., Kitching, T. D., & Verde, L. (2007). On model selection forecasting, dark energy and modified gravity. *Monthly Notices of the Royal Astronomical Society*, *380*(3), 1029-1035.

Hibon, M., Crone, S., & Kourentzes, N. (2012). Statistical Significance of Forecasting Methods. In *32nd Annual International Symposium on Forecasting, Boston, MA, USA*.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, *4*(4), 43-46.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jenkins, A. (2022). Stockouts Defined. *Oracle NetSuite*. https://www.netsuite.com/portal/resource/articles/inventory-management/stockout.shtml

Jiang, A., Tam, K. L., Guo, X., & 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.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, *1*(2), 112-133.

Johnston, F. R., Boylan, J. E., & Shale, E. A. (2003). An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items. *Journal of the Operational Research Society*, *54*(8), 833-837.

Kaya, G. O., Sahin, M., & Demirel, O. F. (2020). Intermittent demand forecasting: a guideline for method selection. *Sādhanā*, *45*(1), 1-7.

Kaya, G. O., Sahin, M., & Demirel, O. F. (2020). Intermittent demand forecasting: A guideline for method selection. Sādhanā, 45, 1-7.

Kerzel, U. (2023). Demand Models For Supermarket Demand Forecasting. *International Journal of Supply and Operations Management*, *10*(1), 89-104.

Kiefer, D., Grimm, F., Bauer, M., & Van Dinther, C. (2021). Demand forecasting intermittent and lumpy time series: Comparing statistical, machine learning and deep learning methods.

Kincheloe, JL 2001, 'Describing the bricolage: conceptualizing a new rigor in qualitative research, Qualitative Inquiry, vol. 7, no. 6, pp. 679–92.

Kincheloe, JL 2005, 'On to the next level: continuing the conceptualization of the bricolage', Qualitative Inquiry, vol. 11, no. 3, pp. 323–50.

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

Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*, *181*, 145-153.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291-302.

Krajewski, L.J., Ritzman, L.P. and Malhotra, M.K. (2012). Operations Management: Processes and Supply Chain. 10th Edition, Prentice Hall, New Jersey.

Laptev, N., Yosinski, J., Li, L. E., & Smyl, S. (2017, August). Time-series extreme event forecasting with neural networks at uber. In *International conference on machine learning* (Vol. 34, pp. 1-5). sn.

Liu, Y., Li, B., Tan, R., Zhu, X., & Wang, Y. (2014). A gradient-boosting approach for filtering de novo mutations in parent–offspring trios. *Bioinformatics*, *30*(13), 1830-1836.

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

Mackay, D. J. C. (1992). *Bayesian methods for adaptive models*. California Institute of Technology.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International journal of forecasting*, *9*(4), 527-529.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. L. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. Journal of Forecasting, 1(2), 111-153. https://doi.org/10.1002/for.3980010202

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54-74.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M5 competition: A major step in the forecasting field. International Journal of Forecasting, 36(1), 1-7. https://doi.org/10.1016/j.ijforecast.2019.09.013

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*.

Mandelbrot, B. B. (1997). The variation of certain speculative prices. In *Fractals and scaling in finance* (pp. 371-418). Springer, New York, NY.

Martin, D., Spitzer, P., & Kühl, N. (2020). A new metric for lumpy and intermittent demand forecasts: Stock-keeping-oriented prediction error costs. *arXiv preprint arXiv:2004.10537*.

McCue, I. (2020). Inventory Carrying Costs: What It Is & How to Calculate It. Oracle NetSuite. https://www.netsuite.com/portal/resource/articles/inventory-management/inventory-carrying-costs.shtml

Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022, December). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach. In *Operations Research Forum* (Vol. 3, No. 4, pp. 1-22). Springer International Publishing.

Mobarakeh, N. A., Shahzad, M. K., Baboli, A., & Tonadre, R. (2017). Improved forecasts for uncertain and unpredictable spare parts demand in business aircraft's with bootstrap method. *IFAC-PapersOnLine*, *50*(1), 15241-15246.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. Journal of Mathematical Psychology, 72, 6-18.

Ng, A., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 14.

Nikolopoulos, K. (2010). Forecasting with quantitative methods: the impact of special events in time series. *Applied Economics*, *42*(8), 947-955.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., & 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.

Nowok, B. (2017). Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control. https://cran.r-project.org/. https://cran.r-project.org/web/packages/synthpop/index.html

Nowok, B., Raab, G. M., & Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 1. *Statistical Journal of the IAOS*, *33*(3), 785-796.

omotive spare parts. *International Journal of Production Economics*, *161*, 1-16.

Pacheco, I., & Barba, R. (2015).Classification and forecasting for Inventory Management of Spare Parts.

Pinçe, Ç., Turrini, L., & Meissner, J. (2021). Intermittent demand forecasting for spare parts: A critical review. *Omega*, *105*, 102513.

PseudoLab. (2020). PseudoLab . 1. Introduction to Time Series - PseudoLab Tutorial Book. Retrieved March 7, 2023, from https://pseudo-lab.github.io/Tutorial-Book-en/chapters/en/time-series/Ch1-Time-Series.html

Ruiz-Aguilar, J. J., Turias, I., Moscoso-López, J. A., Jiménez-Come, M. J., & Cerbán, M. (2016). Forecasting of short-term flow freight congestion: A study case of Algeciras Bay Port (Spain). *Dyna*, *83*(195), 163-172.

Rumelhart, D., Hinton, G., Williams, R., 1988. Parallel Distributed Processing Explorations in the Microstructure of Cognition. MIT Press.

Saeed, A., Li, C., Gan, Z., Xie, Y., & Liu, F. (2022). A simple approach for short-term wind speed interval prediction based on independently recurrent neural networks and error probability distribution. Energy, 238, 122012.

Saini, T. D., Weller, J., & Bridle, S. L. (2004). Revealing the nature of dark energy using Bayesian evidence. *Monthly Notices of the Royal Astronomical Society*, *348*(2), 603-608.

Sak, H., Senior, A., Rao, K., Irsoy, O., Graves, A., Beaufays, F., & Schalkwyk, J. (2015, April). Learning acoustic frame labeling for speech recognition with recurrent neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4280-4284). IEEE.

SAS Institute. (2020). Automatic Intermittent Demand Model Selection. SAS Help Center. Retrieved March 4, 2023, from https://documentation.sas.com/doc/en/hpfug/15.3/hpfug_hpfdet_sect034.htm.

Saxena, D., & Cao, J. (2021). Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR), 54*(3), 1-42.

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. Journal of Big Data, 7(1), 1-22.

Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory management and production planning and scheduling* (Vol. 3, p. 30). New York: Wiley.

Sivia D. S., 1996, Data Analysis: A Bayesian Tutorial. Oxford Univ. Press, Oxford

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 1017-1024).

Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International journal of production economics*, *71*(1-3), 457-466.

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

Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, *26*(1), 134-143.

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

The M5 competition - mofc.unic.ac.cy. (2021). Retrieved March 7, 2023, from https://mofc.unic.ac.cy/wp-content/uploads/2020/02/M5-Competitors-Guide_Final.pdf

Tiwari, M. K., & Chatterjee, C. (2010). Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. *Journal of Hydrology*, *394*(3-4), 458-470.

Trafalis, T. B., & Gilbert, R. C. (2006). Robust classification and regression using support vector machines. European Journal of Operational Research, 173(3), 893-909.

Trinh, H. D., Zeydan, E., Giupponi, L., & Dini, P. (2019). Detecting mobile traffic anomalies through physical control channel fingerprinting: A deep semi-supervised approach. IEEE Access, 7, 152187-152201.

Tsao, Y. C., Kurniati, N., Pujawan, I. N., & Yaqin, A. M. A. (2019, May). Spare Parts Demand Forecasting in Energy Industry: A Stacked Generalization-Based Approach. In *Proceedings of the 2019 International Conference on Management Science and Industrial Engineering* (pp. 163-167).

Van der Auweraer, S., Boute, R. N., & Syntetos, A. A. (2019). Forecasting spare part demand with installed base information: A review. *International Journal of Forecasting*, *35*(1), 181-196.

Waller, D. (2015). Methods for Intermittent Demand Forecasting. Lancaster University. https://www.lancaster.ac.uk/pg/waller/pdfs/Intermittent_Demand_Forecasting.pdf

Walström, P. & Segerstedt, A. (2010). Evaluation of forecasting error measurements and techniques for intermittent demand. International Journal of Production Economics, 128, 625-636.

Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, *20*(3), 375-387.

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

Yin, Z., Chang, K. H., & Zhang, R. (2017, August). Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2131-2139).

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

Zhang, J., Yang, Y., & Ding, J. (2023). Information criteria for model selection. Wiley Interdisciplinary Reviews: Computational Statistics, e1607.

Zhang, W., Luo, Y., Zhang, Y., & Srinivasan, D. (2020). SolarGAN: Multivariate solar data imputation using generative adversarial network. *IEEE Transactions on Sustainable Energy*, *12*(1), 743-746.

Zhuang, X., Yu, Y., & Chen, A. (2022). A combined forecasting method for intermittent demand using the automotive aftermarket data. *Data Science and Management*.

Appendix A: Tables

Table 3: Standalone and Hybrid Approaches for Spare Part Demand Forecasts

Standalone Methods		**Hybrid Methods**		
Method/Technique	Reference	Method/Technique(s)	Method/Technique	Reference
Moving Average **(MA)**	Yule* (1909)	Multiple Aggregation Prediction Algorithm **(MAPA)**	**MA, WMA,** or **ES**	Kourentzes et al. (2013)
Weighted Moving Average **(WMA)**	Holt* (1957)	Recurrent Neural Network **(RNN)**	**SBA**	Fu et al. (2018)
Exponential Smoothing **(ES)**	Brown (1956) *	Elastic Net **(EN)** & Random Forest **(RF)**	**TSB**	Tsao et al. (2019)
Croston's **(CR)**	Croston (1972)	Neural Network **(NN)** & Linear Regression **(LR)**	**CR**	Guitierrez et al. (2008)
Syntetos & Boylan Approximation **(SBA)**	Syntetos & Boylan (2001)	Random Forrest **(RF)** & Linear Regression (**LR)**	**XGB**	Mitra et al. (2022)
Teunter-Syntetos-Babai **(TSB)**	Babai et al. (2019)	Support Vector Machine **(SVM)** Random Forest (**RF)**	**NN**	Dai and Cao* (2017)
Extreme Gradient Boosting **(XGBoost)**	Chen et al. (2015)	UNISON	**RNN, SBA** & **ARIMA**	Wenhan Fu and Chien (2019)
Neural Network **(NN)**	Kourentzes (2013)	ADIDA	**CR**, **TSB** or **SBA**	Nikolopoulos et al. (2011)
Bootstrapping **(BS)**	Willemain (2004)			

Note: The hardware specifications to run the tests are 2.30GHZ processing speed and 32GB RAM.

* Indicates the original author who proposed the method or approach.

Table 4B: Model Complexity Chart

	Methods/Reasons		
MA (5)	Simple averaging technique, limited in capturing complex patterns.	**SBA & RNN (3)**	Combines a simple method (SBA) with a complex one (RNN), increases overall complexity.
WMA (5)	Weighted averaging, slightly more complex than MA, but still simple.	**NN (3)**	Neural Networks, capable of capturing complex patterns, requires more computational resources and can be difficult to interpret.
ES (5)	Exponential smoothing, simple and easy to implement, limited in capturing complex patterns.	**EN, RF & TSB (2)**	Combines Elastic Net, Random Forest, and TSB, which increases complexity by combining multiple models.
CR (5)	Croston's method, focused on intermittent demand, relatively simple calculation.	**NN, LR & CR (2)**	Combines Neural Networks, Linear Regression, and Croston's method, increasing overall complexity.
SBA (5)	Syntetos-Boylan Approximation, an extension of Croston's method, still fairly simple.	**RF, LR & XGB (2)**	Combines Random Forest, Linear Regression, and XGBoost, increasing complexity through the combination of multiple models.
BS (5)	Bootstrapping, a resampling technique with relatively simple calculations, limited in capturing complex patterns.	**SVM & NN (2)**	Combines Support Vector Machine and Neural Networks, both complex techniques, increasing the overall complexity.
TSB (4)	Teunter-Syntetos-Babai, incorporates state space modeling, more complex than CR and SBA.	**ADIDA (1)**	Aggregated Disaggregation and Densification, combines multiple techniques and requires extensive computational resources.
XGB (4)	XGBoost, gradient boosting machine learning algorithm, requires parameter tuning and more computational resources.	**UNISON (1)**	Combines traditional and deep learning techniques (SBA, ARIMA, RNNs), increases complexity by incorporating various methods and multiple layers of data processing.
MAPA (4)	Multiple Aggregation Prediction Algorithm, uses temporal aggregation, more complex than simple averaging methods.		

Table 7: Model Computation Time for each Dataset (minutes)

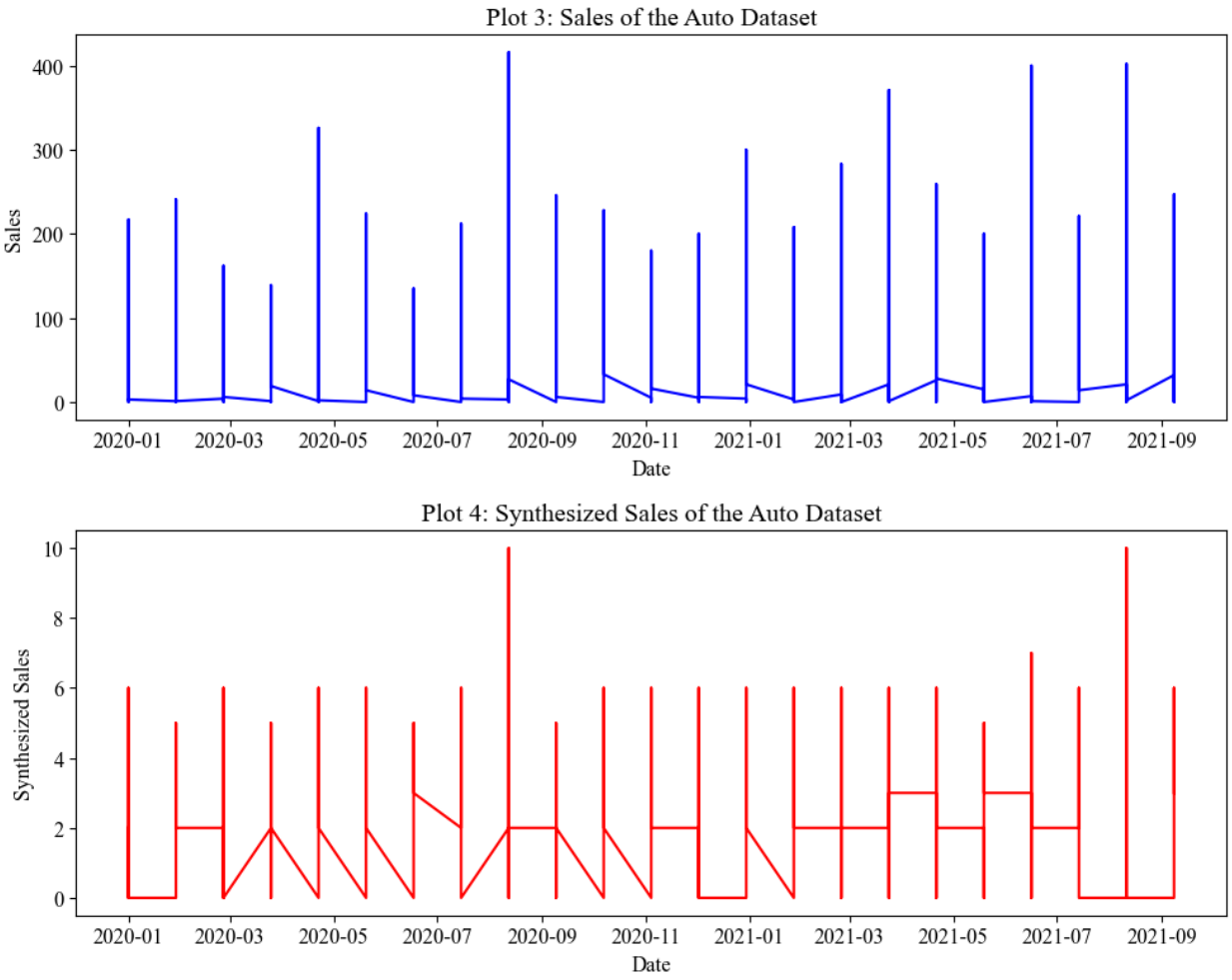
		Datasets				
Type	Method	OIL/SYN	AUTO/SYN	BRAF/SYN	MAN/SYN	ST/SYN
Traditional	**MA**	5.51	1.19	3.88	7.56	13.59
	WMA	17.01	0.43	15.36	14.9	125.96
Non-Traditional (Benchmark)	**ES**	0.02	0.1	0.04	0.04	0.06
	CR	0.24	0.44	0.4	0.54	2.3
	SBA	0.24	0.39	1.45	1.23	2.44
	TSB	0.24	0.43	1.71	0.59	2.28
Machine Learning	**XGB**	0.59	0.14	0.34	1.24	2.24
Deep Learning	**NN**	8.21	1.14	6.32	7.23	25.47
Resampling	**BS**	121.95	21.91	100.93	220.92	420.58
Hybrid Stacking	**RF, LR & XGB**	3.29	0.23	3.34	2.59	46.51
	EN, RF & TSB	3.57	0.57	4.71	3.56	11.62
	NN, LR & CR	8.89	1.58	6.72	8.45	27.77
Non-Linear Combination	**SVM & NN**	560.04	356.56	435.35	476.05	328.58
	SBA & RNN	12.34	2.21	13.24	14.94	24.96
Temporal Aggregation	**MAPA & MA**	0.5	0.03	0.4	0.05	0.08
	MAPA & WMA	0.2	0.01	0.3	0.04	0.06
	MAPA & ES	0.4	0.03	0.6	0.02	0.07
Aggregation Across Series	**ADIDA & CR**	1.23	0.42	2.75	1.09	8.3
	ADIDA & TSB	1.45	0.34	2.41	1.31	7.9
	ADIDA & SBA	1.39	0.42	2.35	1.03	8.09
Temporal Aggregation (Hybrid)	**UNISON**	32.56	3.18	28.56	42.45	1.43
Total computation (Hours):		13.00	6.53	10.52	13.43	17.67

Description: This table presents computation times in minutes for various forecasting methods applied to sales and synthesized sales datasets, including traditional, non-traditional, machine learning, deep learning, and hybrid approaches. The results highlight differences in processing times across methods and datasets.

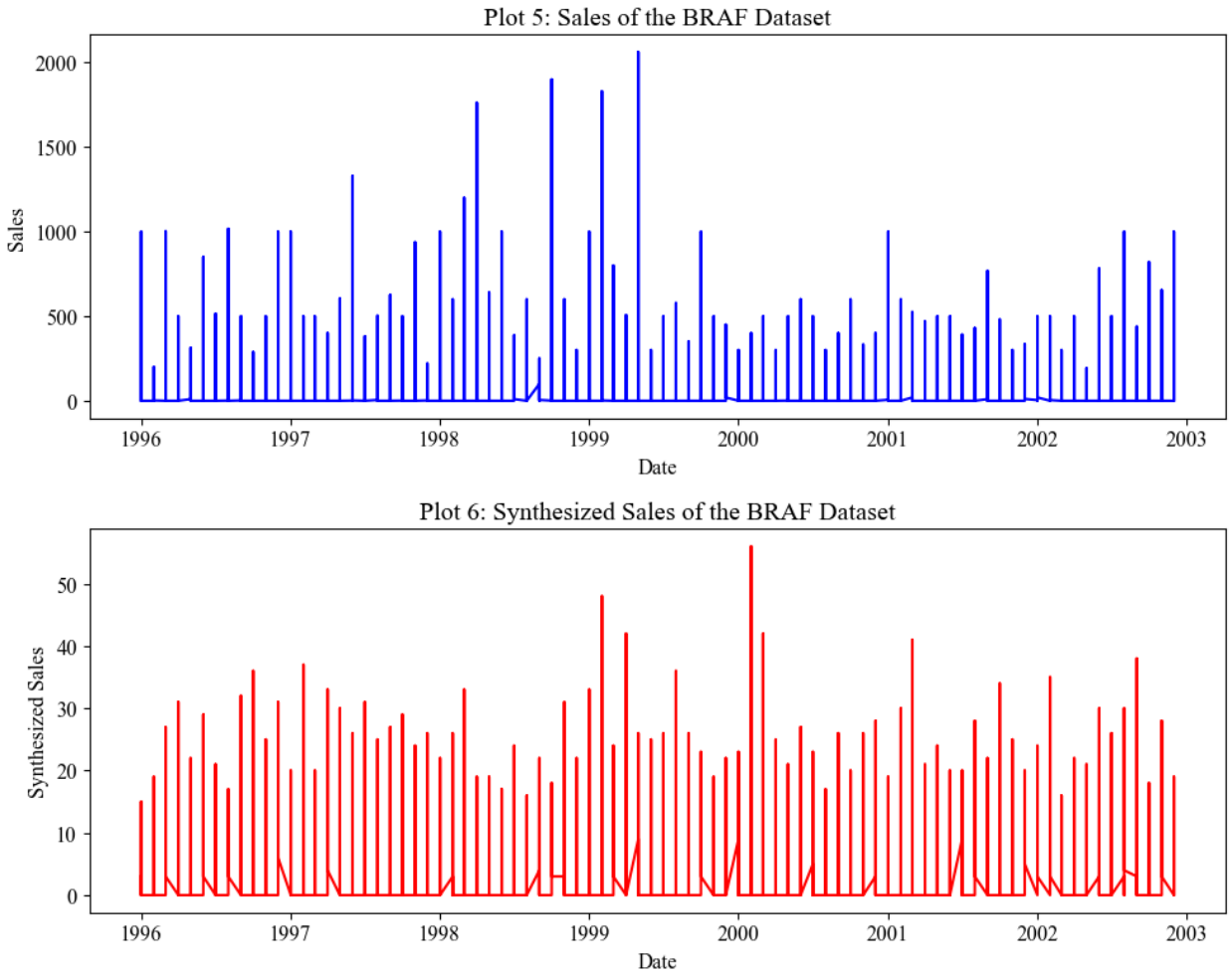
Appendix B: Time Series Plots

Plot 1 & 2: The Time Series Plot of the Actual and Synthesized Sales of the OIL Dataset

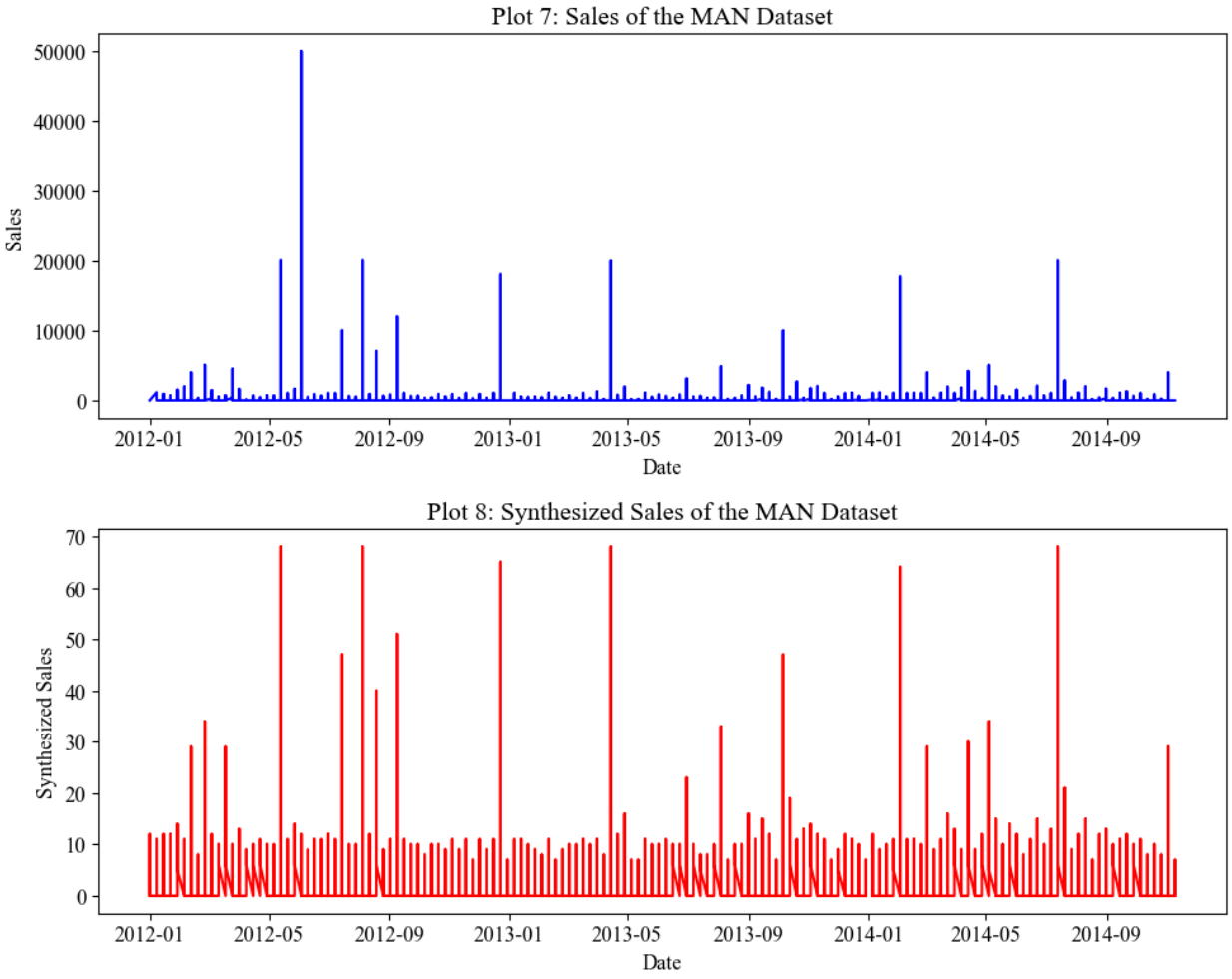
Plot 3 & 4: The Time Series Plot of the Actual and Synthesized Sale of the AUTO Dataset



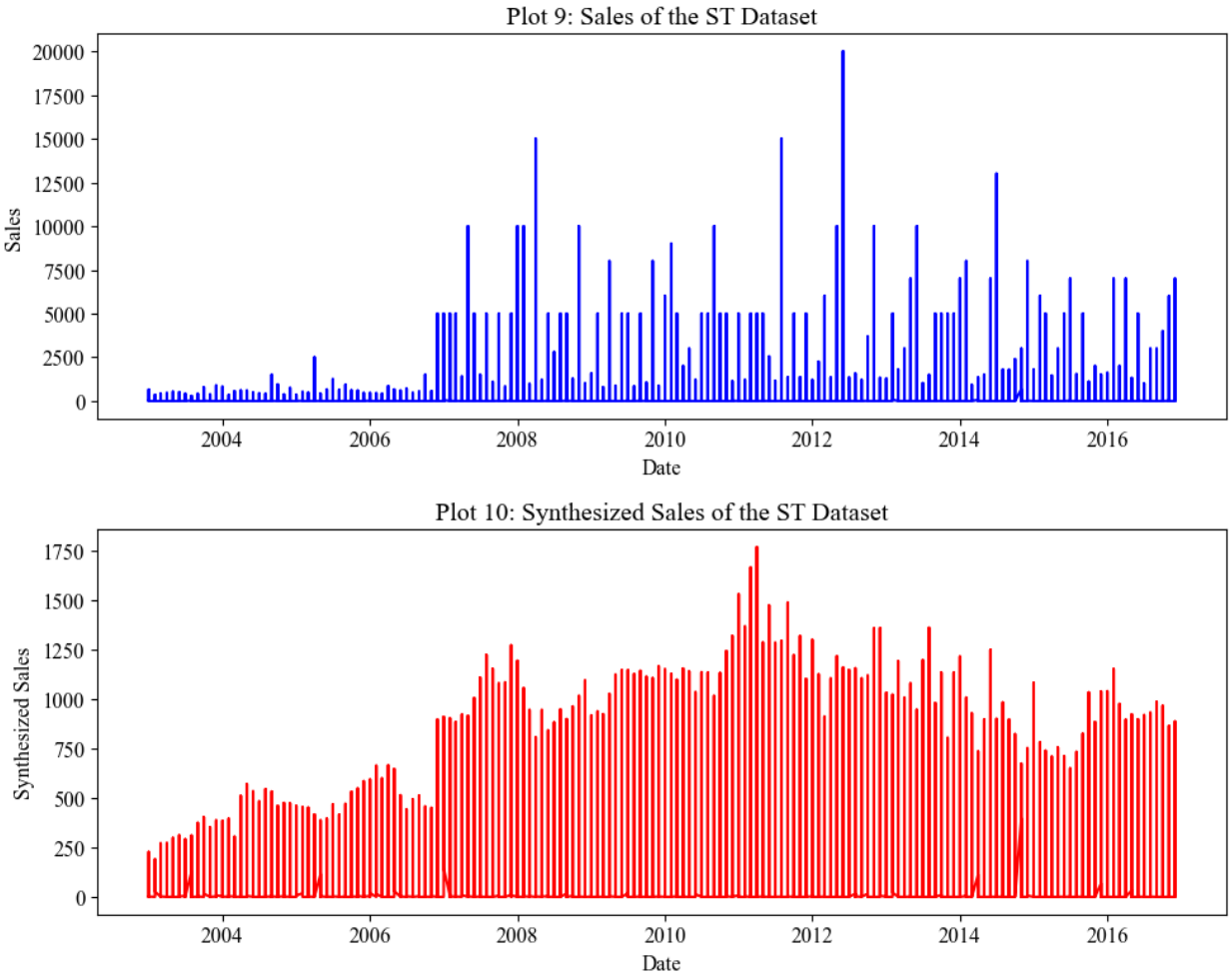
Plot 5 & 6: The Time Series Plot of the Actual and Synthesized Sale of the BRAF Dataset



Plot 7 & 8: The Time Series Plot of the Actual and Synthesized Sale of the MAN Dataset.



Plot 9 & 10: The Time Series Plot of the Actual and Synthesized Sale of the ST Dataset.



Appendix C: Output Tables

Output 1B: Standalone Forecasting Results on Train Data

Method		Metric	Data Set: OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
Traditional Time Series	**MA**	RMSE	10.69	10.55	16.95	125.43	55.28	0.97	0.59	1.56	2.78	20.9
		RMSSE	1	0.99	1.00	1.00	1	1.14	0.89	1.02	1.02	1
		MASE	0.99	0.81	0.94	0.82	0.94	1.91	1.44	0.89	2.23	1.14
		sMAPE	0.96	0.50	0.94	0.97	0.9	0.98	0.09	0.50	0.93	0.87
	WMA	RMSE	11.28	9.23	17.89	125.56	49.37	1.00	0.46	1.67	2.84	42.3
		RMSSE	1.06	0.87	1.06	1.00	0.89	1.18	0.71	1.09	1.04	1
		MASE	1.04	0.71	0.98	0.84	0.83	1.89	0.96	0.96	2.23	0.52
		sMAPE	0.52	0.46	0.69	0.74	0.77	0.13	0.06	0.47	0.93	1
	ES	RMSE	11.14	10.74	17.67	125.53	57.41	0.98	0.54	1.65	2.83	21.54
		RMSSE	1.044	1.01	1.04	1.00	1.04	1.15	0.82	1.08	1.04	1.03
		MASE	1.03	0.83	0.98	0.84	0.97	1.89	1.16	0.95	2.23	0.97
		sMAPE	0.96	0.49	0.97	0.98	0.93	0.42	0.07	0.51	0.93	0.87
Non-Traditional Time Series (Benchmarks)	**CR**	RMSE	11.51	5.72	17.42	125.50	51.93	1.03	0.07	2.16	3.15	19.06
		RMSSE	1.08	0.54	1.03	1.00	0.93	1.21	0.11	1.41	1.15	0.91
		MASE	1.39	0.35	1.08	0.90	1.04	3.20	0.01	1.17	2.25	0.95
		sMAPE	0.98	0.31	0.96	0.98	0.92	0.99	0.00	0.00	0.94	0.85
	SBA	RMSE	11.13	7.09	17.36	125.49	50.87	1.02	0.50	1.77	2.97	18.72
		RMSSE	1.043	0.67	1.02	1.00	0.92	1.20	0.76	1.16	1.09	0.89
		MASE	1.10	0.53	0.96	0.83	0.86	3.16	1.05	0.96	1.99	0.79
		sMAPE	0.98	0.49	0.96	0.98	0.94	0.99	0.06	0.34	0.96	0.88
	TSB	RMSE	11.51	5.72	17.42	113.52	51.93	1.03	0.07	2.16	2.64	19.06
		RMSSE	1.08	0.54	1.03	0.99	0.94	1.21	0.11	1.41	0.97	0.01
		MASE	1.39	0.35	1.08	1.29	1.04	3.20	0.01	1.17	2.22	0.95
		sMAPE	0.98	0.31	0.96	0.96	0.92	0.99	0.00	0.00	0.87	0.85
Machine Learning	**XGB**	RMSE	10.08	7.29	16.58	125.47	39.58	10.67	10.78	16.84	125.47	52.47
		RMSSE	0.94	0.69	0.98	1.00	0.72	1.00	1.02	0.99	1.00	0.95
		MASE	0.96	0.52	0.93	0.83	0.49	0.64	0.68	0.82	0.83	0.52
		sMAPE	0.98	0.54	0.95	0.98	0.91	0.99	0.47	0.96	0.98	0.92
Deep Learning	**NN**	RMSE	13.31	1.67	24.18	126.61	15.78	1.07	0.11	1.88	2.92	4.04
		RMSSE	1.25	0.16	1.43	1.01	0.29	1.26	0.16	1.23	1.07	0.19
		MASE	0.965	0.17	1.02	0.86	0.19	1.79	0.13	1.10	2.42	0.12
		sMAPE	0.96	0.29	0.94	0.99	0.91	0.97	0.01	0.41	0.92	0.79
Resampling Techniques	**BS**	RMSE	10.67	10.61	16.95	125.43	10.61	0.85	0.65	1.89	3.13	0.65
		RMSSE	1	1.00	1.00	1.00	1	1.00	1.00	1.23	1.15	1
		MASE	0.99	0.82	0.94	0.82	0.82	1.87	1.70	1.00	1.63	1.7
		sMAPE	0.96	0.51	0.94	0.97	0.51	1.00	0.10	1.00	0.92	0.1
Granularity:			Monthly	Monthly	Monthly	Weekly	Monthly	Monthly	Monthly	Monthly	Weekly	Monthly
# of Observations (Thousands)			160	18	125	362	1242	160	18	125	362	1242

***Description**:* Output 1B shows the forecasting performance of different models on the train data of five data sets and its synthesized counterpart.: OIL, AUTO, BRAF, MAN, and ST. The metrics used to evaluate the models include RMSE, RMSSE, MASE, and sMAPE. The models are classified into two types: traditional and non-traditional. The traditional models include MA, WMA, ES, and XGB, while the non-traditional models include CR, SBA, and TSB. In general, the traditional models perform better than the non-traditional models, with XGB being the best performer for most products. The worst performer is the SBA model, which has high errors across all metrics for most products. The table provides insights into the strengths and weaknesses of the different forecasting models in fitting the train data.

Output 2B: Combined Forecasting Results on Train Data

Method(s)		Accuracy Metric	Data Set									
			OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
Hybrid Stacking	**SBA & RNN**	RMSE	10.79	13.21	19.32	125.78	62.44	1.02	0.81	1.77	2.97	23.08
		RMSSE	1.01	1.24	1.14	1.00	1.13	1.20	1.24	1.16	1.09	1.10
		MASE	1.04	0.95	0.97	0.83	1.03	3.16	1.76	0.96	1.99	0.97
		sMAPE	0.97	0.35	0.94	0.99	0.87	0.99	0.06	0.34	0.96	0.88
	EN, RF & TSB	RMSE	10.67	0.50	16.95	114.99	10.50	0.85	0.65	1.53	2.73	19.84
		RMSSE	1.00	0.05	1.00	1.00	0.19	1.00	0.99	1.00	1.00	0.95
		MASE	0.99	0.03	0.94	1.03	0.43	1.87	1.69	0.87	2.46	0.83
		sMAPE	0.96	0.22	0.94	0.97	0.91	1.00	0.10	0.51	0.90	0.82
	NN, LR & CR	RMSE	10.67	1.54	16.95	114.99	55.28	0.85	0.04	1.53	2.73	20.95
		RMSSE	1.00	0.14	1.00	1.00	1.00	1.00	0.07	1.00	1.00	1.00
		MASE	0.99	0.16	0.94	1.03	0.94	1.87	0.07	0.87	2.46	0.92
		sMAPE	0.96	0.31	0.94	0.97	0.90	1.00	0.00	0.51	0.90	0.83
	RF, LR & XGB	RMSE	10.07	7.29	16.57	114.99	39.55	10.67	10.77	16.83	114.99	52.47
		RMSSE	0.94	0.69	0.98	1.00	0.72	1.00	1.02	0.99	1.00	0.95
		MASE	0.96	0.52	0.93	1.03	0.49	0.65	0.68	0.82	0.83	0.52
		sMAPE	0.98	0.54	0.96	0.97	0.91	0.99	0.48	0.96	0.97	0.92
Non-Lienar Combination	**SVM & NN**	RMSE	10.67	14.54	16.95	125.43	16.17	0.85	0.88	1.53	2.73	9.58
		RMSSE	1.00	1.37	1.00	1.00	0.68	1.00	1.34	1.00	1.00	0.47
		MASE	0.97	1.07	0.96	1.08	0.29	1.94	1.78	0.88	2.45	0.31
		sMAPE	0.96	0.35	0.94	0.97	0.88	1.00	0.01	0.50	0.91	0.69
Granularity:			Monthly	Monthly	Monthly	Weekly	Monthly	Monthly	Monthly	Monthly	Weekly	Monthly
# of Observations (Thousands)			160	18	125	362	1242	160	18	125	362	1242

Description: Output 2B shows the results of combined forecasting on the train dataset using different methods and models. The columns represent various evaluation metrics such as RMSE, RMSSE, MASE, and sMAPE, and the rows represent different models and methods used for forecasting. The first set of rows presents the results of hybrid stacking using SBA & RNN, while the second set of rows shows the results of hybrid stacking using EN, RF & TSB. The third set of rows shows the results of hybrid stacking using NN, LR & CR, while the fourth set of rows present the results of non-linear combination using RF, LR & XGB. The last set of rows presents the results of non-linear combination using SVM & NN.

Output 3B: Aggregate Forecasting Results on Train Data

Method(s)		Accuracy Metric	Data Set									
			OIL	AUTO	BRAF	MAN	ST	OIL SYN	AUTO SYN	BRAF SYN	MAN SYN	ST SYN
Temporal Aggregation	**MAPA & MA**	RMSE	1419	5197	1554	14543	7052	349.93	2415	154	1554	3191
		RMSSE	0.99	1.00	1.00	0.98	0.57	0.98	1.00	0.90	0.60	0.35
		MASE	0.73	0.85	0.80	0.64	1.24	0.78	1.00	0.79	0.34	2.40
		sMAPE	0.11	0.10	0.09	0.15	0.16	0.13	0.09	0.01	0.03	0.08
	MAPA & WMA	RMSE	1767	6970	1665	16214	13168	469	3249	200	3156	9299
		RMSSE	1.23	1.34	1.07	1.09	1.07	1.31	1.34	1.16	1.22	1.03
		MASE	0.85	1.16	0.83	0.67	2.25	1.04	1.29	1.00	0.73	7.19
		sMAPE	0.13	0.14	0.09	0.16	0.27	0.17	0.11	0.01	0.06	0.20
	MAPA & ES	RMSE	1419	5525	1554	14543	7052	350	2573	154	1554	3191
		RMSSE	0.99	1.06	1.00	0.98	0.57	0.98	1.06	0.90	0.60	0.35
		MASE	0.73	0.94	0.80	0.64	1.24	0.78	1.14	0.79	0.34	2.40
		sMAPE	0.11	0.11	0.09	0.15	0.16	0.13	0.10	0.01	0.03	0.08
Aggregation Across Series	**ADIDA & CR**	RMSE	31.50	9.56	51.58	243	122.85	0.91	0.49	1.51	2.50	18.51
		RMSSE	2.95	0.90	3.04	2.12	2.22	1.06	0.75	0.99	0.92	0.88
		MASE	7.49	0.69	7.09	5.84	4.63	2.80	0.78	0.77	2.01	0.90
		sMAPE	0.96	0.40	0.95	0.91	0.91	0.99	0.02	0.31	0.85	0.86
	ADIDA & TSB	RMSE	23.18	9.12	36.40	167.38	84.86	0.91	0.49	1.51	2.50	18.51
		RMSSE	2.17	0.86	2.15	1.46	1.54	1.06	0.75	0.99	0.92	0.88
		MASE	6.02	0.63	5.25	3.86	3.26	2.80	0.78	0.77	2.01	0.90
		sMAPE	0.96	0.47	0.95	0.92	0.91	0.99	0.02	0.31	0.85	0.86
	ADIDA & SBA	RMSE	31.50	9.56	51.58	243.38	122.85	0.90	0.50	1.39	2.24	18.50
		RMSSE	2.95	0.90	3.04	2.12	2.22	1.06	0.76	0.91	0.82	0.88
		MASE	7.49	0.69	7.09	5.84	4.63	2.74	1.04	0.73	1.90	0.75
		sMAPE	0.96	0.40	0.95	0.91	0.91	0.99	0.07	0.45	0.87	0.90
Temporal Aggregation (Hybrid)	**UNISON**	RMSE	14.10	4.40	16.84	153.47	42.30	16.02	18.90	23.66	162.37	87.30
		RMSSE	0.88	0.24	0.69	0.94	0.44	1.00	1.02	0.97	1.00	0.91
		MASE	0.79	0.25	0.65	0.71	0.44	0.59	0.60	0.68	0.66	0.53
		sMAPE	0.92	0.50	0.89	0.95	0.85	0.96	0.38	0.90	0.96	0.86
Granularity:			Monthly	Monthly	Monthly	Weekly	Monthly	Monthly	Monthly	Monthly	Weekly	Monthly

Description: Output 3B shows aggregate forecasting results for the train datasets using the methods and models motivated in the literature section. The table includes five datasets, and the columns with "SYN" indicate the synthesized counterparts. The evaluation metrics used to assess the forecasting accuracy vary across the different models and methods.

Output 5: ***Model Selection Criteria Composite Scores***

Method	Metric	ST	OIL	MAN	BRAF	AUTO	Complexity Tier	Average Performance	Normalized Performance	Composite Score
ADIDA-TSB	MASE	3.35	2.01	4.21	5.24	0.70	1	3.10	1	0.80
ADIDA-CR	MASE	3.35	2.01	4.21	5.24	0.70	1	3.10	1	0.80
ADIDA-CR	MASE	2.42	1.91	2.87	3.92	0.66	1	2.36	0.71	0.57
SVM-NN	MASE	8.79	2.44	1.11	1.00	0.92	2	2.85	0.90	0.54
SVM	MASE	8.26	2.20	0.77	0.85	0.96	2	2.61	0.81	0.48
RF-EN-TSB	MASE	4.09	2.55	1.10	0.90	0.95	2	1.92	0.53	0.32
MAPA WMA	MASE	0.70	0.62	3.80	0.48	1.38	2	1.40	0.33	0.20
MAPA ES	MASE	1.10	0.85	1.18	0.79	2.62	2	1.31	0.30	0.18
CR-LR-NN	MASE	1.09	2.45	1.04	0.98	0.83	2	1.28	0.28	0.17
MAPA MA	MASE	1.38	0.50	1.09	0.82	2.23	2	1.20	0.25	0.15
RF-LR-XGB	MASE	0.54	2.45	1.04	0.97	0.54	2	1.11	0.22	0.13
NN	MASE	0.81	2.38	0.80	0.87	0.91	3	1.16	0.24	0.09
RNN	MASE	0.94	2.25	0.78	0.84	0.94	3	1.15	0.23	0.09
SBA-RNN	MASE	0.90	1.96	0.77	0.83	0.73	3	1.04	0.19	0.08
TSB	MASE	1.02	2.21	1.15	0.93	0.46	4	1.15	0.23	0.05
XGB	MASE	0.53	2.47	0.86	0.97	0.54	4	1.08	0.20	0.04
UNISON	MASE	0.49	0.66	0.69	0.65	0.35	1	0.57	0.00	0.00
WMA	MASE	0.87	2.19	0.78	0.86	0.86	5	1.11	0.22	0
RF	MASE	0.54	0.02	0.38	0.90	0.95	2	0.56	0	0
BS	MASE	1.04	1.62	0.84	0.98	0.83	5	1.06	0.20	0
SBA	MASE	0.87	2.20	0.77	0.84	0.60	5	1.06	0.20	0
CR	MASE	1.02	2.20	0.83	0.93	0.95	5	1.19	0.25	0
ES	MASE	0.90	2.20	0.78	0.86	0.84	5	1.12	0.22	0
MA	MASE	8.79	2.22	0.84	0.98	0.83	5	2.73	0.85	0

Description: This table presents the performance of the motivated forecasting methods using the average K-fold CV Mean Absolute Scaled Error (MASE) loss across five different datasets: ST, OIL, MAN, BRAF, and AUTO. Each method is assigned a Complexity Tier, with Tier 1 being the most complex and Tier 5 the least. The table also includes average performance, normalized performance, and a composite score for each method. The composite score combines normalized performance and complexity, aiming to identify simple models with good performance. ADIDA-TSB and ADIDA-CR, both with Complexity Tier 1, have the highest composite scores (0.8) despite their complexity, indicating strong performance. As the composite score decreases, models with lower complexity tiers are observed, highlighting the trade-off between complexity and performance. For example, methods with Complexity Tier 5, such as WMA, RF, BS, SBA, CR, and ES, have lower composite scores. This suggests that their performance is not as strong relative to their simplicity. Decision-makers should consider this balance when selecting forecasting methods for their specific tasks.