UNIVERSITY OF ICELAND

Magister Scientiarum thesis
in Industrial Engineering

A systematic categorization process facilitating the selection of demand forecasting methods

Júlíus Ingi Guðmundsson

May 2022

FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

A systematic categorization process facilitating the selection of demand forecasting methods

Júlíus Ingi Guðmundsson

Magister Scientiarum thesis in Industrial Engineering

Tutor: Gunnar Stefánsson, professor

Secondary tutor: Fanndís Kristinsdóttir

Examiner: Burkni Helgason

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science

School of Engineering and Natural Sciences, University of Iceland

May 2022

A systematic categorization process facilitating the selection of demand forecasting methods

This thesis satisfies 30 credits towards an M.Sc.
in Industrial Engineering in the Faculty of Industrial Engineering, Mechanical Engineering and Computer Science,
University of Iceland, School of Engineering and Natural Sciences

Abstract

Demand forecasting is the process of estimating the demand for products or services in future time periods. Forecasts are usually generated for different stock keeping units (SKUs) and selecting the appropriate forecasting methods for multiple SKUs can be challenging. Numerous categorizations have been proposed to facilitate the selection of forecasting methods by grouping SKUs with similar characteristics together. The SBC categorization scheme has received significant attention in literature and categorizes SKUs according to their underlying demand patterns. The purpose of this study was to compare several categorizations and identify the categorizations that led to the lowest forecasting error for each distinct demand pattern. The additional categorizations were comprised of a trend analysis, seasonality detection and ABC analysis. The forecasts were generated with Croston's method, the Syntetos-Boylan Approximation, the Teunter-Syntetos-Babai method and random forest. The forecasting performances were then compared based on the mean absolute scaled error (MASE) and root mean square error (RMSE) accuracy measures. Most of the SKUs in a data set from ÁTVR were found to have a smooth demand pattern. Approximately half as many SKUs had intermittent or lumpy demand patterns and the fewest number of SKUs had erratic demand patterns. Various combinations of categorizations were tested and the results generally showed that the trend analysis led to the lowest MASE value for intermittent and lumpy demand patterns while the ABC analysis produced lower MASE values for smooth and erratic demand patterns. Furthermore, the ABC analysis led to the lowest RMSE value for all the demand patterns.

Útdráttur

Eftirspurnarspár eru notaðar til að meta framtíðareftirspurn á vörum eða þjónustu. Spár eru oftast gerðar fyrir ólík vörunúmer og það getur verið krefjandi að velja viðeigandi spáaðferðir fyrir mikinn fjölda af vörunúmerum. Margar flokkunaraðferðir hafa í gegnum tíðina verið notaðar til að einfalda val á spáaðferðum með því að flokka saman vörunúmer með svipaða eiginleika. SBC flokkunarkerfið hefur fengið mikla athygli fræðimanna, en það flokkar vörunúmer eftir undirliggjandi eftirspurnarmynstrum. Tilgangur verkefnisins var að bera saman nokkrar flokkunaraðferðir og greina hvaða flokkanir leiddu til lægstu spávillunnar fyrir hvert og eitt eftirspurnarmynstur. Flokkunaraðferðirnar sem voru prófaðar greindu þróun á breytingu eftirspurnar, hvort að árstíðarbundnar sveiflur væru til staðar og viðeigandi flokk úr ABC greiningu. Framtíðareftirspurn hvers vörunúmers var spáð með Croston aðferðinni, Syntetos og Boylan nálguninni, Teunter-Syntetos-Babai aðferðinni og slembiskógi. Frammistaða aðferðanna var svo borin saman með sköluðu meðaltölugildisfráviki (MASE) og staðalfráviki leifa (RMSE) nákvæmnimælikvörðunum. Flest vörunúmer í gagnasetti frá ÁTVR reyndust hafa slétt eftirspurnarmynstur. Um það bil helmingi færri vörunúmer voru með slitrótt eða skrykkjótt eftirspurnarmynstur og fæstur fjöldi vörunúmera var með óreglulegt mynstur. Ýmsar samsetningar af flokkunum voru prófaðar og niðurstöðurnar sýndu að flokkunin fyrir þróun á breytingu eftirspurnar leiddi almennt til lægstu MASE gildanna fyrir slitróttu og skrykkjóttu eftirspurnarmynstrin en ABC greiningin til lægstu MASE gildanna fyrir sléttu og óreglulegu eftirspurnarmynstrin. ABC greiningin leiddi einnig til lægstu RMSE gildanna fyrir öll eftirspurnarmynstrin.

Table of Contents

List of Figures

List of Tables

Abbreviations

ADI – Average demand interval

ÁTVR – The State Alcohol and Tobacco Company of Iceland

CRO – Croston's method

CV^2 – Squared coefficient of variation

MAE – Mean absolute error

MASE – Mean absolute scaled error

ML – Machine learning

MSE – Mean square error

RF – Random forest

RMSE – Root mean square error

SBA – Syntetos-Boylan Approximation

SBC – Syntetos, Boylan and Croston

SKU – Stock keeping unit

TSB – Teunter-Syntetos-Babai

Acknowledgements

Many wonderful people assisted me in one way or another throughout the writing process and I am thankful for the support I received. I would like to thank my supervisor Gunnar Stefánsson for giving me valuable feedback and helping me formulate the research questions.

The project was done in collaboration with AGR Dynamics and The State Alcohol and Tobacco Company of Iceland. I am grateful for the data they provided and their willingness to respond to my various questions. I would also particularly like to thank Fanndís Kristinsdóttir, technical consultant at AGR Dynamics, for insightful discussions and for introducing me to the critical issue that the thesis sought to address.

Finally, I would like to thank my family and friends for supporting me throughout my studies. My parents have always been there for me and for that I am eternally grateful.

1 Introduction

1.1 Background

For centuries, shop owners and businesses have relied on experience and instinct when planning and ordering inventory for upcoming periods. The objective for decision-makers is to ensure their customers' demands are met by having the right products available at the right time, while also minimizing the cost of buying and holding inventory (Thomopoulos, 2015). It is therefore essential for businesses to reliably forecast future demand and reduce the overall waste in their operation (Lolli et al., 2017). The consequences of poor forecasting can be detrimental, as it may cause an over- or undersupply of inventory. An oversupply of inventory can lead to higher inventory holding costs and higher product obsolescence while an undersupply of inventory can lead to lost sales and a damaged reputation (Kharfan et al., 2021; Thomopoulos, 2015). Demand forecasting serves as the basis for planning and inventory management in organizations and, in fact, it sets the entire supply chain in motion (Boylan & Syntetos, 2021).

A multitude of different forecasting methods are available today and the most appropriate method for a given scenario will depend on the circumstances. Forecasting methods can be chosen manually but it becomes increasingly more difficult to reliably forecast demand as complexity and the scale of the operation increases. Organizations can have a huge number of stock keeping units (SKUs) and the process of finding suitable forecasting methods for all individual SKUs can be very time consuming (Fildes et al., 2008). Sophisticated software is therefore commonly used to automate the selection process. The first step usually involves eliminating inappropriate forecasting methods from consideration because it is very computationally expensive for the software to test all forecasting methods for every single SKU in an organization (Spiliotis et al., 2020). Deciding which forecasting methods to exclude is not always clear-cut and the decisions can be justified based on various conditions and arbitrary values. The forecasting software handles these decisions, among many other things, but it is often perceived as a black box due to its complexity. It is, however, crucial for decision-makers to understand the set of principles in the internal workings of the system so that they can trust the forecasts that the software generates (Chase, 2021).

The performance and the applicability of the software is largely dependent on the validity of the forecasting method selection process. If relevant forecasting methods are excluded, then the likelihood of suboptimal forecasting results increases. Contrarily, if too many forecasting methods are tested, then it will take longer to wait for the results because the run time increases. Additionally, arbitrary values and conditions that are suitable for one organization do not necessarily work well for another organization. A systematic approach that is generally applicable and supported by both theoretical and empirical research is consequently preferred. Implementing a framework that categorizes SKUs so that only appropriate forecasting methods are considered in the selection process is therefore of great significance (Boylan et al., 2008).

A universally applicable categorization framework has not been developed, likely due to the lack of consensus about the appropriateness of the possible categorizations. SKUs can be categorized in numerous ways and the categorizations are typically based on historical sales data. The amount sold, the standard deviation of the sales, the average period between demands and the average demand are examples of information that can be retrieved from historical sales data to categorize SKUs. The Syntetos, Boylan and Croston (SBC) categorization scheme is a well-established approach that utilizes historical sales data to categorize the underlying demand patterns of SKUs. It uses the average period between non-zero demands and the variation of non-zero demand sizes to categorize SKUs as "intermittent", "lumpy", "smooth" or "erratic" (Syntetos et al., 2005). Other common categorizations include the ABC analysis, measuring the strength or presence of seasonality, detecting if there are positive or negative trends present and detecting if there are short-term or long-term trends present (Boylan & Syntetos, 2021; Hyndman & Athanasopoulos, 2021).

A major limitation of using historical sales data to categorize SKUs is that sufficient periods of existing data are required for the process. New products do not have any previous sales history and traditional forecasting methods are consequently not appropriate to use until enough historical data has been collected (Ching-Chin et al., 2010). There are, however, ways to circumvent this. If the product is a newer version of an existing product, then the sales history of the older product can be linked to the new product to create a fake sales history. Historical sales data can also be created by clustering similar products together with machine learning and estimating the demand from products that are already established (Kharfan et al., 2021). Seasonal products are another example of products that can be difficult to forecast accurately with historical sales data, especially products that have periods of zero demand. The data is very sparse when products are only in demand for certain periods of time each year, which increases the likelihood of the forecast having high uncertainty (Boylan & Syntetos, 2021).

The need for a transparent and effective categorization process is becoming increasingly apparent for demand forecasting. Multiple categorizations have been extensively researched but the effectiveness of combining different categorizations into one unified process has received limited academic interest (Bandeira et al., 2020). Increasing the number of categorizations will lead to categories with fewer SKUs but each category can tune the forecasting method parameters differently, possibly achieving better forecasting performance in the process. Determining which categorizations contribute the most to reducing the forecasting error and finding the ideal number of categorizations is critical so that organizations can improve their forecasting performance.

1.2 Purpose

The purpose of this research project is to develop an effective categorization process that facilitates the selection of appropriate demand forecasting methods. The process aims to identify the underlying demand patterns of SKUs and subsequently to further categorize the SKUs in a data set with relevant alternative approaches. The forecasting performance of various combinations of categorizations are analysed for each demand pattern and the top performing categorizations are highlighted. The process also provides clarity for the choice of forecasting method at any given time for each SKU because demand planning software can be perceived as a black box due to its complexity. Fundamentally, the process strives to enhance forecasting performance by using historical sales data to group SKUs with similar characteristics together and generating forecasts tailored to each group.

The proposed research questions are as follows:

- How many SKUs in the data set were categorized as having each distinct underlying demand pattern according to the SBC categorization scheme?
- Which categorization or combination of categorizations led to the lowest demand forecasting error for each distinct underlying demand pattern?

1.3 Scope

The focus of this research is to implement a generalizable categorization process and identify the structure that offers the greatest demand forecasting accuracy. The categorizations are generated with historical sales data because it is readily available to organizations and it also enables the process to be applicable in various scenarios. Different combinations of categorizations are tested and compared solely based on the selected forecasting accuracy measures. Other potentially relevant factors for organizations such as logistics and distribution, inventory stock levels, inventory holding costs, marketing promotions and safety stocks are consequently not within the scope of this research.

1.4 Thesis outline

This report has been structured in six main chapters. The first chapter introduces the topic of demand forecasting and highlights the importance of systematically categorizing products. A literature review is then conducted to evaluate what researchers have published on the topic. The third chapter describes the methodology used in this study. Next, the results are presented and they are then subsequently discussed in the discussion chapter. Finally, the results are summarized and the research questions are answered in the conclusion chapter.

2 Literature review

2.1 Demand forecasting

Demand forecasting is the process of estimating the demand for products or services in future time periods. It is critical for organizations to plan and control activities effectively and demand forecasting enables them to take informed decisions about future operations (Boylan & Syntetos, 2021). Many variables can affect future demand and the uncertainty of the circumstances make it very challenging to produce accurate forecasts. The predictions are consequently rarely perfect but forecasting methods and processes can still be improved to provide better and more reliable results (Krajewski et al., 2019).

The future demand for some products are more difficult to forecast than others and studies have highlighted the importance of developing specialized methods to handle them more fittingly (Croston, 1972; Nikolopoulos, 2021; Syntetos & Boylan, 2005). Products with intermittent demand are, for instance, difficult to forecast because they experience periods of zero demand (Boylan & Syntetos, 2021). New products are also notoriously difficult to forecast as they can have very limited or no historical data (Ching-Chin et al., 2010). As a result, numerous forecasting methods have been developed over the years and they can be relevant in diverse conditions. Researchers are constantly seeking ways to improve demand forecasting performances and they have, due to the sheer number of options, suggested several systematic ways to select the most appropriate forecasting methods (Thomopoulos, 2015).

2.2 Demand pattern categorization

A common categorization approach is to categorize products based on their underlying demand pattern. Williams (1984) laid the foundation by developing a conceptual categorization scheme, capable of categorizing products whose demand is sporadic and with periods of zero demand. His approach is based on variance partition, where the variance of demand during a lead time is split into three separate components: transaction variability, demand size variability and lead-time variability. These components are then used to categorize product demand as “smooth”, “slow-moving” or “sporadic”. Table 2.1 presents Williams’ categorization scheme.

Table 2.1: Demand categorization scheme by Williams (1984).

Lead-time demand component			
Transaction variability	**Demand size variability**	**Lead-time variability**	**Demand pattern categorization**
Low			Smooth
High	Low		Slow moving
High	High	Low	Sporadic
High	High	High	Sporadic, with highly variable lead time

Note. Adapted from "Stock Control with Sporadic and Slow-Moving Demand" by T. Williams, 1984, *Journal of the Operational Research Society*, 35(10), p. 947. Copyright 1984 by the Operational Research Society Ltd.

The categorization scheme has, however, been criticized for its lack of generalizability because the boundaries for the categories are chosen subjectively by management (Boylan et al., 2008; Syntetos et al., 2005). Eaves and Kingsman (2004) analysed Williams' categorization scheme with data from the Royal Air Force and they concluded that the categorization scheme does not adequately categorize demand patterns. They argued that smooth demand should be distinguished by both transaction variability and demand size variability instead of only considering transaction variability. Eaves and Kingsman consequently proposed a revised version of the categorization scheme that can be seen in Table 2.2.

Table 2.2: Demand categorization scheme by Eaves and Kingsman (2004).

Lead-time demand component			
Transaction variability	**Demand size variability**	**Lead-time variability**	**Demand pattern categorization**
Low	Low		Smooth
Low	High		Irregular
High	Low		Slow moving
High	High	Low	Mildly intermittent
High	High	High	Highly intermittent

Note. Adapted from "Forecasting for the ordering and stock-holding of spare parts" by A. H. Eaves and B. G. Kingsman, 2004, *Journal of the Operational Research Society*, 55(4), p. 432. Copyright 2004 by the Operational Research Society Ltd.

The category of "irregular" demand was added to the categorization scheme to account for demand size variability when the transaction variability is low. The word "sporadic" was also replaced with the word "intermittent" (Eaves & Kingsman, 2004). The reason for changing the word was not specified but the words are often used interchangeably in literature (Nikolopoulos, 2021; Syntetos et al., 2011). The revised categorization scheme has faced the same criticism as the original scheme by Williams because management must subjectively choose the boundaries for the categories (Boylan et al., 2008; Syntetos et al., 2005).

Syntetos et al. (2005) suggested a new and more broadly applicable categorization approach, named the SBC categorization scheme, where lead times are assumed to be constant. Their motivation was to replace arbitrary or subjective cut-off values with values supported by theoretical and empirical research. To achieve this, they approached the categorization process in a different way than the other categorization schemes did (Boylan et al., 2008). The conventional approach was to start the categorization process by choosing arbitrary boundaries for the demand pattern categories and then to analyse which methods are suitable for each category. Instead, they compared forecasting methods based on theoretically quantified error measures to find the boundaries of superior performance and they then used the results to define the demand patterns (Syntetos et al., 2005).

Two parameters, the average demand interval (ADI) and the squared coefficient of variation (CV^2), are used to divide demand patterns into four quadrants (Syntetos et al., 2005), as shown in Figure 2.1. The ADI is sometimes denoted by p in literature and measures the average period between demands by dividing the total number of periods with the number of non-zero demand periods. Products with multiple periods of zero demand will therefore have higher ADI values than products with relatively stable demands (Placencia et al., 2021). The CV^2, on the other hand, measures the variation of the demand sizes in non-zero periods by dividing the standard deviation of the sample with the average value of the sample and then squaring the result. Essentially, the ADI indicates the level of intermittency while CV^2 indicates the level of erraticness (Spiliotis et al., 2020).

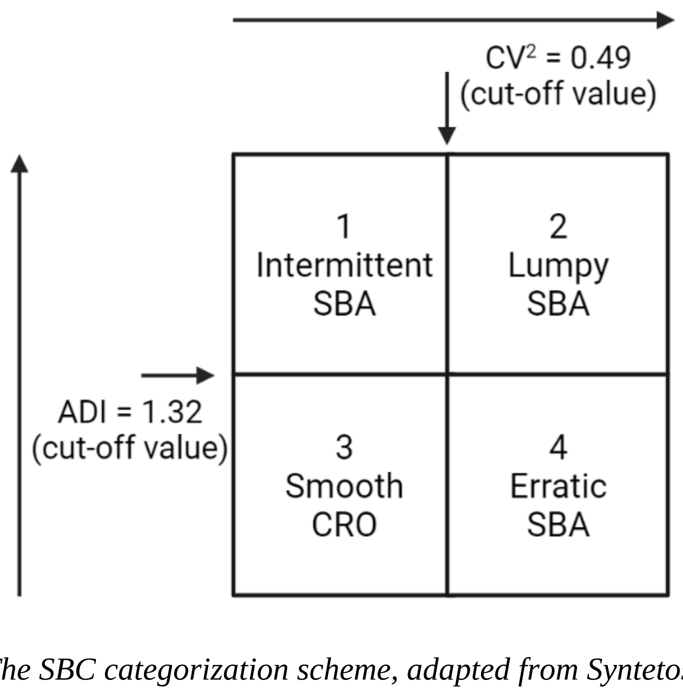

Figure 2.1: The SBC categorization scheme, adapted from Syntetos et al. (2005).

Croston (1972) was the first to suggest using ADI, along with the average demand size, to estimate underlying demand patterns and he proposed a forecasting method specifically designed to forecast intermittent demand. Johnston and Boylan (1996) later compared

Croston's method (CRO) with a commonly used method for short-term forecasting called simple exponential smoothing (SES), which only considers demand size. Their goal was to determine the conditions when standard forecasting methods are no longer suitable to forecast intermittent demand. They concluded that CRO demonstrated superior performance when the ADI is greater than 1.25 and the performance continues to improve as the ADI increases. Syntetos and Boylan (2001) later showed that CRO is biased and they proposed a modified version of CRO, named the Syntetos-Boylan Approximation (SBA), that reduces the bias with a correction factor (Syntetos & Boylan, 2005). Next, Syntetos et al. (2005) compared CRO, SES and SBA to find the boundaries of superior forecasting performance and they subsequently determined the demand pattern categorization rules for ADI and CV^2. They tested the methods both theoretically with mean squared error (MSE) as the accuracy measure and empirically on 3,000 products with intermittent demand in the automotive industry to validate the results. The cut-off values defining the four demand pattern quadrants in the SBC categorization scheme, which they label as "intermittent", "lumpy", "smooth" and "erratic", were determined to be ADI = 1.32 and $CV^2 = 0.49$.

Subsequent studies have questioned both the validity of the SBC categorization scheme cut-off values and the accuracy measures used to evaluate the forecasting utility of the categorization scheme, which are presented in Chapter 2.5 (Heinecke et al., 2013). The cut-off values are approximations because they are derived from inequalities describing the relationship between CRO and SBA. Kostenko and Hyndman (2006) were first to point out that the inequalities could be simplified to define the boundary more accurately between the superior performance of CRO and SBA. Heinecke et al. (2013) and Petropoulos and Kourentzes (2015) later conducted empirical analyses that supported the simplifications. The smooth demand quadrant of the SBC categorization scheme is the only quadrant that is affected by the simplifications and the suggested adjustments can be seen in Figure 2.2. Additionally, Boylan et al. (2008) performed a case study on roughly 16,000 SKUs to estimate the usefulness of having a hard-coded ADI cut-off value when identifying intermittence and they concluded that cut-off values ranging from 1.18-1.86 periods can be applicable, suggesting that organizations can benefit from adjusting the cut-off value for their operation instead of relying on a predetermined cut-off value.

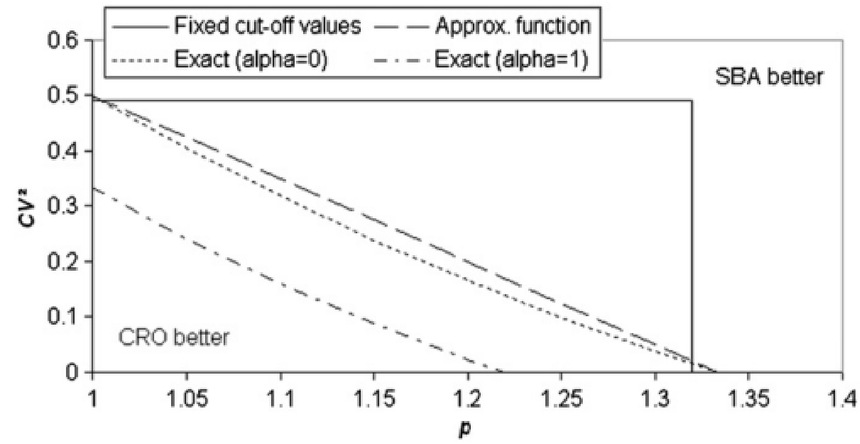

Figure 2.2: Disputed parameter space in the SBC categorization scheme, from Heinecke et al. (2013).

Despite the criticism, the SBC categorization scheme is frequently cited in literature and has been used in numerous case studies and experiments (Boylan et al., 2008; Costantino et al., 2018; Doszyn, 2020; Gelei & Dobos, 2020; Liu, 2020; Placencia et al., 2021; Wallström & Segerstedt, 2010). Recent studies have emphasized comparing the performances of various machine learning forecasting methods against statistical methods for some, or all, of the SBC categorization scheme demand patterns (de Oliveira et al., 2020; Gutierrez et al., 2008; Lolli et al., 2017; Rožanec et al., 2021; Spiliotis et al., 2020, 2021; Tsao et al., 2019). Interestingly, very few studies have included the adjustments suggested by Kostenko and Hyndman (2006). A likely explanation is that the demand patterns in question already have relatively low ADI and CV^2 values, which generally makes them easier to forecast accurately than demand patterns with higher values (Regattieri et al., 2005). Newer studies are also not constrained to only examining the original three forecasting methods. Researchers usually test multiple forecasting methods when utilizing the SBC categorization scheme and do not see the benefit of applying additional rules that are based on unrelated forecasting methods. A logical next step would therefore be to extend the original theoretical MSE analysis of the SBC categorization scheme to other, more recent, forecasting methods so that the most appropriate method is recommended for each category (Heinecke et al., 2013).

Linkages between the SBC categorization scheme and demand distributional assumptions for SKUs have also been investigated in literature (Turrini & Meissner, 2019). Syntetos et al. (2011) analysed the performance of various demand distributions with data from three empirical databases and proposed a classification rule that is based on the parameters from the SBC categorization scheme, shown in Figure 2.3. The implications are important for inventory management because the performance of inventory policies can be improved by selecting the most fitting demand distributions for SKUs (Turrini & Meissner, 2019).

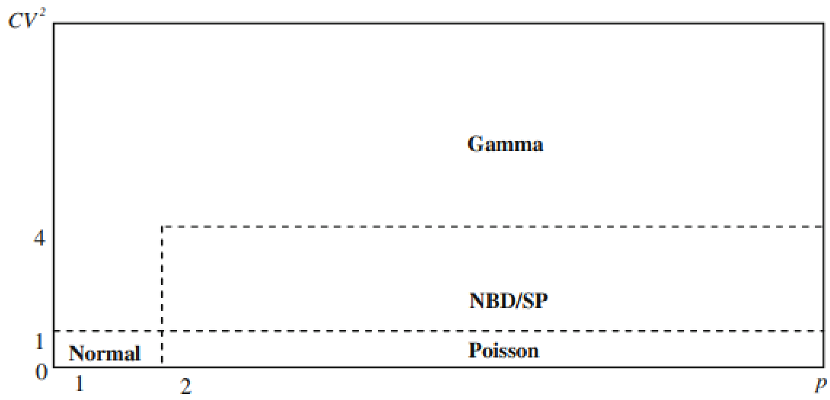

Figure 2.3: Classification rule for demand distributional assumptions, from Syntetos et al. (2011).

Other variations of demand pattern categorizations have been researched but none have garnered the same amount of interest as the SBC categorization scheme. Kiefer at al. (2021) implemented a variation of the categorization scheme proposed by Williams (1984), where only the transaction variability and demand size variability were used to categorize demand patterns. Varghese and Rossetti (2008) noted that most categorization schemes in literature only consider lumpiness and intermittence. They suggested considering an additional parameter, dependence, ranging from high negative correlation to high positive correlation when categorizing demand patterns. The performance of their proposed categorization scheme has not been compared to other categorization schemes, but they did, however, show in their preliminary study with real data from the US Navy that forecasting error was reduced, when compared with the existing process.

2.3 Alternative categorizations of SKUs

Demand pattern categorization provides one way of grouping SKUs with similar characteristics together, but it is certainly not the only way to categorize SKUs. A handful of notable categorizations are discussed in the following sections.

2.3.1 Trend analysis

Trend analysis is commonly used to detect the presence or absence of trends in time series data (Brillinger, 1994). On the surface, it seems straightforward to define and identify trend but there is disagreement among researchers about the precise definition. Some consider trend the part of a time series that changes slowly over time while others regard it as a general direction of future long-term movements (Harvey, 2016). Ultimately, trend is a vague concept that is contingent on the circumstances and several statistical tests are available, classified as either parametric, semi-parametric or non-parametric, to estimate the impact of trends in time series data (Brillinger, 1994; Hamed & Rao, 1998).

Two widely used statistical trend tests are the t-test for detecting linear trend and the Mann-Kendall trend test (Gauthier & Hawley, 2015; Rutkowska, 2015). Both methods test the null hypothesis (H_0) of no trend being present against the alternative hypothesis (H_a) of trend being present. The t-test is a parametric test, where the underlying distribution is assumed to fit a normal distribution, and that reduces the applicability of the test because underlying distributions are not necessarily normally distributed (Gauthier & Hawley, 2015). The Mann-Kendall trend test (Kendall, 1955; Mann, 1945), on the other hand, is a non-parametric test, meaning that it works for all distributions. Furthermore, trends do not have to be linear and researchers have shown that the test is slightly stronger than the Cox-Stuart test, which is another comparable non-parametric trend test (Rutkowska, 2015). Despite the test being relatively robust, it still requires independent and identically distributed samples. The data must therefore have no autocorrelation and no seasonal effects for the Mann-Kendall trend test to perform adequately (Hirsch & Slack, 1984).

Researchers have proposed several modified versions of the Mann-Kendall trend test to address the shortcomings of the original test (Hussain & Mahmud, 2019). Both Hamed and Rao (1998) and Yue and Wang (2004) have addressed the autocorrelation problem by suggesting a variance correction approach. Hirsch et al. (1982) proposed a version of the Mann-Kendall trend test for seasonal time series, where the seasonal trend is calculated. The tests require different amounts of samples to function properly but all the tests generally perform better with large amounts of samples (Hamed & Rao, 1998). Researchers should therefore exercise caution before utilizing the trend tests if they aim to use only recent samples to estimate short-term trends. It is up to researchers to assess the circumstances and choose appropriate trend tests that suit their data.

2.3.2 Seasonality detection

Seasonality in time series refers to periodic fluctuations that occur at specific regular interval lengths, such as weekly, monthly or yearly (Brockwell & Davis, 2016). It is essential for decision-makers when detecting the presence of seasonality in time series to have a clear purpose of how the information will be utilized because numerous approaches are available, that usually operate under different assumptions. This will reduce the likelihood of utilizing inappropriate methods and generating misleading results. If the purpose is to check whether a time series is non-stationary, then the Augmented Dickey Fuller test is commonly used (Patterson, 2012). Another possibility is to use an autocorrelation plot and the Ljung-Box test to identify non-stationary time series by either plotting the time series directly or decomposing it so that the seasonal component can be plotted separately. Non-stationary behaviour can, however, be caused by several different factors, so it is inappropriate to draw the conclusion that all non-stationary time series show seasonal behaviour (Hyndman & Athanasopoulos, 2021). If the goal is to examine whether multiple time series show seasonal behaviour without assuming their distribution, then the Kruskal-Wallis test is a suitable option (Ostertagová et al., 2014).

Kruskal and Wallis (1952) proposed a non-parametric test that determines whether there are statistically significant differences between the medians of two or more groups of samples. Their test is known as the Kruskal-Wallis test and, like the Mann-Kendall trend test in Chapter 2.3.1, it is a statistical test that uses hypothesis testing to draw conclusions about the data. It works by ranking the samples of a group and then comparing them with other groups to check if at least one sample statistically dominates another sample. The groups are defined as specific intervals, often for each month of the year, and only contain samples from that period. The output of the test does, however, not indicate which groups are statistically significantly different or for how many pairs of groups it occurs, it only states whether at least one group statistically dominates another.

The Kruskal-Wallis test is well-suited for identifying time series with seasonality and a benefit of the test is that the number of samples in each group do not have to be equal (Ostertagová et al., 2014). The test can therefore be used throughout the year instead of waiting for all weeks, months or other specific intervals of the year to have the same number of samples. The main drawback of the test is, however, that misleading results can be generated if the number of samples in each group is very small. It is consequently recommended to have at least five samples in each group (Kruskal & Wallis, 1952; Gauthier & Hawley, 2015).

2.3.3 ABC analysis

ABC analysis is a well-established categorization method that divides inventory into three categories based on their perceived importance (Ravinder & Misra, 2014). Organizations might be dealing with thousands of SKUs and the purpose of the categorization is to provide them with a way of highlighting important SKUs (Teunter et al., 2010). The “A” category contains the most important SKUs to a business and decision-makers place high emphasis on managing them closely. The “B” category is moderately important, while the “C” category contains the least important SKUs. The ABC analysis is based on the Pareto principle, where 20% of the SKUs account for 80% of the consumption, but the percentages are seldom precisely accurate, so they are adjusted accordingly. The split for the categories is traditionally set as 20%/80% for “A”, 30%/15% for “B” and 50%/5% for “C” (Boylan & Syntetos, 2021). Essentially, the “A” category has few SKUs with high consumption, there are slightly more SKUs in the “B” category with lower consumption, while the “C” category has many SKUs with little consumption.

The ABC analysis is typically performed by ranking all SKUs based on a criterion that decision-makers deem appropriate for their organization (Douissa & Jabeur, 2016). Only one criterion is used in a traditional ABC analysis and the two most common criteria are demand volume and demand value (Boylan & Syntetos, 2021). The demand volume is determined by simply summing up the number of units sold in each period while the demand value is calculated by multiplying the price of the SKU with the demand volume in each period (Teunter et al., 2010). Other criteria are still important to businesses and notable examples of lesser-known criteria for the ABC analysis include lead time, substitutability, criticality, inventory cost, durability and ordering cost (Douissa & Jabeur, 2016; Rezaei & Dowlatshahi, 2010). Researchers have consequently proposed numerous methods that can utilize multi-criteria for the ABC analysis (Ravinder & Misra, 2014). The categorization process becomes more complex as the number of criteria included increases and comparisons of the perceived importance between SKUs tend to become fuzzier. Ultimately, decision-makers decide what is most important to their organization and the criterion or criteria for the ABC analysis is chosen accordingly (Rezaei & Dowlatshahi, 2010).

2.4 Selection of forecasting methods

Numerous forecasting methods have been developed over the years and decision-makers face the difficult task of selecting appropriate forecasting methods from a wide variety of options (Rožanec et al., 2021). The selection process depends on many factors and studies have shown that there is no single forecasting method that outperforms all other methods in all circumstances (Gelei & Dobos, 2020; Petropoulos & Kourentzes, 2015). As a result, it is critical to identify the strengths and weakness of each method so that they can be applied in the appropriate circumstances (Makridakis et al., 2018). The categorizations presented in Chapters 2.2 and 2.3 simplify the selection process by providing users with systematic ways of grouping products together based on certain criteria. Most studies have, however, limited their scope to only one categorization process and the effects of combining multiple categorizations have not received significant attention in academic literature (Bacchetti & Saccani, 2012). Additionally, certain groups of products are often excluded from examination because they are inappropriate for the categorization schemes. New, phased-out or obsolete products are all examples of groups that usually require special attention before any forecasting method can be considered applicable (Boylan et al., 2008; Teunter et al., 2011).

Forecasting methods are often categorized in academic literature into two broad categories: statistical methods and machine learning (ML) methods (Januschowski et al., 2020). They share many similarities and strive to improve forecasting performance by minimising the error for various accuracy measures (Makridakis et al., 2018). Moreover, their forecasting performance favours small values of ADI and CV^2, indicating that the smooth category is the simplest demand pattern category to forecast accurately (Regattieri et al., 2005). The methods also clearly have their differences and operate under different assumptions (Makridakis et al., 2018).

2.4.1 Statistical methods

Statistical methods have been thoroughly researched and have been used to forecast demand for decades. They function by using a predefined relationship with certain assumptions to forecast the demand with historical data (Vandeput, 2021). It is therefore imperative to select forecasting methods with assumptions that correspond to the products in the data set, otherwise the forecasting performance is likely to suffer. The behaviour of statistical methods is often highly predictable, due to the predefined relationships, and they tend to have more explanatory power than ML methods (Spiliotis et al., 2020). They also have relatively low computational costs, but Januschowski et al. (2020) pointed out that other costs should be considered as well when comparing methods. Statistical methods might ultimately incur higher costs than ML methods because they can require more supervision and tuning from experts to operate effectively.

Croston's method (CRO), Syntetos–Boylan Approximation (SBA) and Teunter-Syntetos-Babai (TSB) method are all examples of well-established statistical forecasting methods. They have all performed well in various circumstances for the SBC categorization scheme and continue to be included in recent studies (Doszyn, 2020; Spiliotis et al., 2020). Descriptions of the notations used for the methods can be seen in Table 2.3.

Table 2.3: Symbol descriptions for the statistical forecasting methods, adapted from Teunter et al. (2011).

Symbol	Description
y_t	Demand in period t
y'_t	Estimate of the demand made in period t for period $t+1$
z_t	Demand size in period t
z'_t	Estimate of the demand size in period t
p_t	Demand interval in period t
p'_t	Estimate of the demand interval in period t
d'_t	Estimate of the demand occurrence probability in period t
α	Smoothing parameter $(0 \leq \alpha \leq 1)$
β	Smoothing parameter $(0 \leq \beta \leq 1)$

CRO was proposed by Croston (1972) and is commonly used to forecast intermittent demand. The estimate of the future demand for each product is produced by dividing the estimated demand size with the estimated demand interval. The demand size and demand interval are then only updated in periods with non-zero demand. Equation 2.1 shows how forecasts are generated with CRO.

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

where $\begin{cases} If\ y_t > 0\text{: } z'_t = z'_{t-1} + \alpha(z_t - z'_{t-1})\ and\ p'_t = p'_{t-1} + \alpha(p_t - p'_{t-1}) \\ Otherwise\text{: } z'_t = z'_{t-1}\ and\ p'_t = p'_{t-1} \end{cases}$.

Syntetos and Boylan (2005) later proposed adding a correctional factor to CRO to reduce the bias and named it the SBA. It has been compared to CRO in multiple studies involving the SBC categorization scheme and shown to commonly produce more accurate forecasts for the intermittent, lumpy and erratic demand patterns (Doszyn, 2020; Spiliotis et al., 2020; Syntetos et al., 2005). Equation 2.2 shows how forecasts are generated with the SBA.

$$y'_t = \left(1 - \frac{\alpha}{2}\right)\frac{z'_t}{p'_t} \tag{2.2}$$

where $\begin{cases} If\ y_t > 0\text{: } z'_t = z'_{t-1} + \alpha(z_t - z'_{t-1})\ and\ p'_t = p'_{t-1} + \alpha(p_t - p'_{t-1}) \\ Otherwise\text{: } z'_t = z'_{t-1}\ and\ p'_t = p'_{t-1} \end{cases}$.

Teunter et al. (2011) proposed another modified version of CRO, known as the TSB method. They emphasized estimating the demand occurrence probability rather than the demand interval. Instead of only updating the demand interval in periods of non-zero demand, like CRO and SBA, the TSB method updates the estimate of the probability of occurrence in every period. The TSB method has been shown to perform well for products that quickly become obsolete (Babai et al., 2019). Equation 2.3 shows how forecasts are generated with the TSB method.

$$y'_t = d'_t z'_t \tag{2.3}$$

where $\begin{cases} If\ y_t > 0\text{: } z'_t = z'_{t-1} + \alpha(z_t - z'_{t-1})\ and\ d'_t = d'_{t-1} + \beta(1 - d'_{t-1}) \\ Otherwise\text{: } z'_t = z'_{t-1}\ and\ d'_t = d'_{t-1} + \beta(0 - d'_{t-1}) \end{cases}$.

2.4.2 Machine learning methods

ML methods, on the other hand, do not assume any a priori relationship and instead learn relationships directly with historical data (Vandeput, 2021). They can use whatever data is available and new information can even be extracted with feature engineering, but variables are eventually prioritized and only a subset of significant variables is chosen to reduce the amount of irrelevant data (Kharfan et al., 2021). Furthermore, ML methods are generally more computationally expensive than statistical methods (Makridakis et al., 2018). Generalizations about the interpretability of ML methods are inappropriate because it varies greatly between methods. For instance, neural networks are usually hard to interpret while decision trees are easy to interpret (Januschowski et al., 2020). The performance of ML methods is consistently getting better and they have been shown to outperform statistical methods in some cases by producing less biased and more accurate forecasts (de Oliveira et al., 2020; Lolli et al., 2017; Spiliotis et al., 2020). Conversely, they have also been found to perform worse than well-established statistical methods (Makridakis et al., 2018). The top performing forecasting methods today combine multiple different methods together, either or both statistical and ML methods, to increase forecasting accuracy (Bandeira et al., 2020; Petropoulos & Kourentzes, 2015; Tsao et al., 2019).

Random forest (RF) is a popular ML method that is used for both classification and regression. It is an ensemble learning method that builds multiple independent decision trees with training data. Each tree then generates a prediction and the average prediction of the trees is used as the final prediction for regression tasks (Breiman, 2001). Variance is reduced when training the decision trees by utilizing the bagging technique. Bagging involves choosing samples of data randomly with replacement, meaning that samples can be chosen more than once, and the trees are therefore able to grow differently (Vandeput, 2021). This process increases the robustness of the predictions and RF has consequently been shown to produce relatively accurate results (Spiliotis et al., 2020; Tsao et al., 2019).

2.5 Evaluation of forecasting performance

Organizations strive to operate at the highest level and face many challenging decisions that can impact their performance. One such critical decision is determining how to evaluate forecasting performance (Boylan & Syntetos, 2021). There are numerous accuracy measures available but there is no single indicator that is suitable for all situations. The accuracy measures have different characteristics and must therefore be carefully chosen to support the goals of the organization (Vandeput, 2021). Davydenko and Fildes (2013) highlighted the importance of considering interpretability, robustness to outliers and scale independence when choosing accuracy measures. Organizations might, for instance, prefer an accuracy measure prioritizing high bias that consistently overshoots the forecasted demand instead of fluctuating forecasts with a lower magnitude of error, when the consequences of a stock-out are severe (Vandeput, 2021). Furthermore, it is common to choose more than one accuracy measure because a single measure is usually unable to capture the full extent of the forecasting error (Lolli et al., 2017; Wallström & Segerstedt, 2010).

Hyndman and Koehler (2006) investigated various measures of forecasting accuracy and identified four distinct categories. They found that the accuracy measures can be based on scale-dependent measures, percentage errors, relative errors and measures, and scaled errors. They also found that many of the measures are not generally applicable and inappropriate use of them will, in fact, generate misleading results. It is therefore important to use measures that are both comprehensible and possess good statistical properties (Koutsandreas et al., 2022). Subsequent studies and reviews of accuracy measures have included additional measures in their analyses and the results suggest similar selection recommendations. Importantly, they all fundamentally agree on the importance of selecting appropriate accuracy measures to compare forecasting methods and note that a perfect measure does not exist (Davydenko & Fildes, 2013; Koutsandreas et al., 2022; Shcherbakov et al., 2013).

2.5.1 Scale-dependent measures

As the name suggests, scale-dependent measures are on the same scale as the data that is being analysed (Hyndman & Athanasopoulos, 2021). The measures are also highly impacted by outliers and comparisons can therefore not be made between multiple data sets with different scales or magnitudes (Shcherbakov et al., 2013). The influence of the outliers can be reduced by cleaning the data or by employing variations of measures that are less sensitive to outliers (Koutsandreas et al., 2022). The advantages of scale-dependent measures are, however, that they are both easy to interpret and simple to calculate (Hyndman & Athanasopoulos, 2021). The mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) are the most popular scale-dependent measures and have been used in numerous studies to compare the performance of various demand forecasting methods (Bandeira et al., 2020; Placencia et al., 2021; Teunter et al., 2011). Equation 2.4 shows how RMSE is calculated, where e_i is the forecast error and N is the number of samples. Overall, the measures are considered very good in the appropriate setting but useless when comparing data with multiple scales (Vandeput, 2021).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i)^2} \tag{2.4}$$

2.5.2 Percentage errors

The accuracy is measured as a percentage and is calculated, in its simplest form, by dividing the forecast error with the actual demand (Hyndman & Koehler, 2006). The mean absolute percentage error (MAPE) is the most well-known and commonly used percentage error accuracy measure (Vandeput, 2021). Variations of percentage error measures involve, for instance, taking the median absolute value or root mean square value of the forecast error. The advantage of using percentage errors is that they are scale-independent (Hyndman & Koehler, 2006). The disadvantages, on the other hand, are that they are biased towards low forecasts and are unable to handle intermittent data with periods of zero demand. Calculations yield infinite or undefined results because the forecasting error is divided by zero (Hyndman & Koehler, 2006; Shcherbakov et al., 2013). The disadvantages far outweigh the advantages of using percentage errors and they are generally considered poor accuracy indicators (Vandeput, 2021).

2.5.3 Relative errors and measures

Relative errors are determined by dividing each error with the error obtained from a benchmark forecasting method. Any method can be chosen as the benchmark but the naïve method, where the next forecast is simply set as the last observation, is usually chosen (Shcherbakov et al., 2013). Similarly, relative measures are calculated by dividing each measure with the measure from a defined benchmark method. The main advantages of relative errors and measures are that they are scale-independent and the results are easy to interpret (Hyndman & Koehler, 2006). Contrarily, the shortcomings include sensitivity towards outliers and undefined results when the error or measure from the benchmark method is equal to zero (Davydenko & Fildes, 2013). Mean relative absolute error (MRAE), geometric mean relative absolute error (GMRAE) and relative mean absolute error (RelMAE) are examples of well-known measures based on relative errors or measures (Hyndman & Koehler, 2006). Additionally, several researchers have highlighted that the relative geometric root mean square error (RelGRMSE) possesses desirable statistical properties that limit the impact of outliers (Davydenko & Fildes, 2013; Eaves & Kingsman, 2004; Heinecke et al., 2013; Syntetos & Boylan, 2005).

2.5.4 Scaled errors

Hyndman and Koehler (2006) proposed scaled errors as a new approach in hopes of avoiding the shortcomings of the other accuracy measures. The fundamental first step in their alternative approach is to scale the errors, as shown in Equation 2.5.

$$q_t = \frac{|e_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|} \tag{2.5}$$

Where e_t denotes the forecast error and the denominator denotes the in-sample MAE from the benchmark naïve method. The mean absolute scaled error (MASE) is then determined by simply calculating the mean of the scaled errors (Hyndman & Koehler, 2006).

$$MASE = mean(|q_t|) \tag{2.6}$$

The advantages of scaled errors include that they are scale-independent, highly unlikely to face the division by zero problem and they are easy to interpret. A value higher than one indicates that the in-sample naïve method produced better results than the forecasting method in question, so lower values of MASE are clearly preferred when comparing forecasting methods (Heinecke et al., 2013). The disadvantages are that the results are undefined when all of the observations are the same and the results can also be severely impacted by extreme cases, such as when dividing by small values that can appear in series with only a few observations (Davydenko & Fildes, 2013; Hyndman & Koehler, 2006). Scaled errors have received a lot of attention in literature and the MASE accuracy measure has been used in many studies (Bandeira et al., 2020; Doszyn, 2020; Heinecke et al., 2013; Kiefer et al., 2021; Tsao et al., 2019).

3 Methodology

This chapter describes the methodology and research design that was used to analyse different combinations of categorization methods and how the performances of various forecasting methods were compared. The methodological approach that was used in this research was a case study with quantitative data analysis. The case study was done in collaboration with both AGR Dynamics and ÁTVR. Case studies are performed by analysing the real-world application of relevant theoretical concepts. A common concern about case studies is that it can be difficult to draw generalized conclusions from the results, so further analysis is often required before the results can be established (Yin, 2018).

3.1 Preparation

A non-systematic literature review was performed to analyse existing research papers, journals, and books on demand forecasting, categorizations, demand patterns, statistical forecasting methods and ML forecasting methods. Only peer-reviewed journals and publications found on licensed database sites were considered sources of valid references. Database sites such as Web of Science, ScienceDirect, ResearchGate and Google Scholar were all used to search for publications relevant to this case study. Once the initial search on the database sites had yielded a starting set of publications, the snowballing approach was used to identify additional publications by investigating the reference lists of the publications from the starting set (Wohlin, 2014).

3.2 Data gathering

The primary data was collected directly from the data source at ÁTVR. The data contained three different types of files: historical sales data, item information and feature information. The historical sales data spanned over a three-year period, from 2017-2019, and was comprised of the daily sales volume for each SKU. The item information file provided details about the SKUs, such as the weight, sale price, whether it was active, etc. Finally, the feature information contained categorical details about the SKUs, such as the manufacturer, packaging container, product group, sale price category, etc.

The literature review, covered in Chapter 2, provided a foundation of existing research that was used to design the categorization process. It also helped identify what information was needed for the analysis and additional data was later requested to ensure the functionality of the categorization process. The additional data included the first recorded sale date of each SKU to help identify new products. It also included the sale category, where the SKUs had been labelled according to their profit margin and availability throughout the year.

3.3 Data cleansing

Multiple steps were taken in the data cleansing process to ensure the data was adequately prepared for the data analysis. The data cleansing, and the subsequent data analysis, was coded in Jupyter Notebook with the Python programming language. The data set included 76,567 SKUs with at least one period of historical sales data and the data cleansing process began by removing the SKUs with incomplete data. SKUs that did not have historical sales data, item information and feature information were removed. This was done to make sure the SKUs had the same type of information available and so that the ML method could be applied without any missing data. Next, SKUs that were labelled in the data as special orders, showcase products or other types of unconventional products that require special attention to forecast were removed. Additionally, inactive SKUs were removed because they were outside the scope of this study. Product returns and inventory implications were also outside the scope of this study, so all instances of negative daily demand were replaced with zero demand.

Furthermore, the data needed to be prepared for the categorization process to avoid distorting the results. The standard deviations and average values are only supposed to be calculated from non-zero demand when determining the CV^2 for the SBC categorization scheme. Consequently, for the remaining SKUs, all periods of zero demand were removed to prepare the data for the categorization process. The final step involved aggregating the data set to a weekly level for the trend analysis and the forecasting process, monthly level for the seasonality detection and yearly level for detecting phased-out and obsolete products. After the data cleansing process, the number of SKUs had been reduced to 6,167.

3.4 Data analysis

Once the data cleansing was completed, the data set was split into a training set with two-thirds of the data and a test set with one-third of the data. Sales data from the years 2017 and 2018 was used as a training set and sales data from 2019 was used as a test set. The years 2020 and 2021 were not included in the case study due to the decline of economic activities during the COVID-19 pandemic. Only the training set was used in the categorization process to identify the characteristics of each product. The forecasting process was then subsequently implemented to test different combinations of the categorizations with several forecasting methods. The results were then compared with two accuracy measures to determine which categorization or combination of categorizations displayed superior performance.

3.4.1 Categorization process

The proposed categorization process was designed based on prominent demand forecasting literature and can be seen in Figure 3.1. Each step in the process requires a specific amount of new data to be updated and it is up to decision-makers to decide how frequently to perform the whole categorization process. The demand for products is often fluctuating and it is very likely that some categories will change as more data is included in the training set. Frequent updates can therefore make comparisons less transparent because products are constantly changing categories and the number of samples in each category varies. The performance was analysed over a one-year period and to simplify interpreting the results, the categorization process was performed once for this case study.

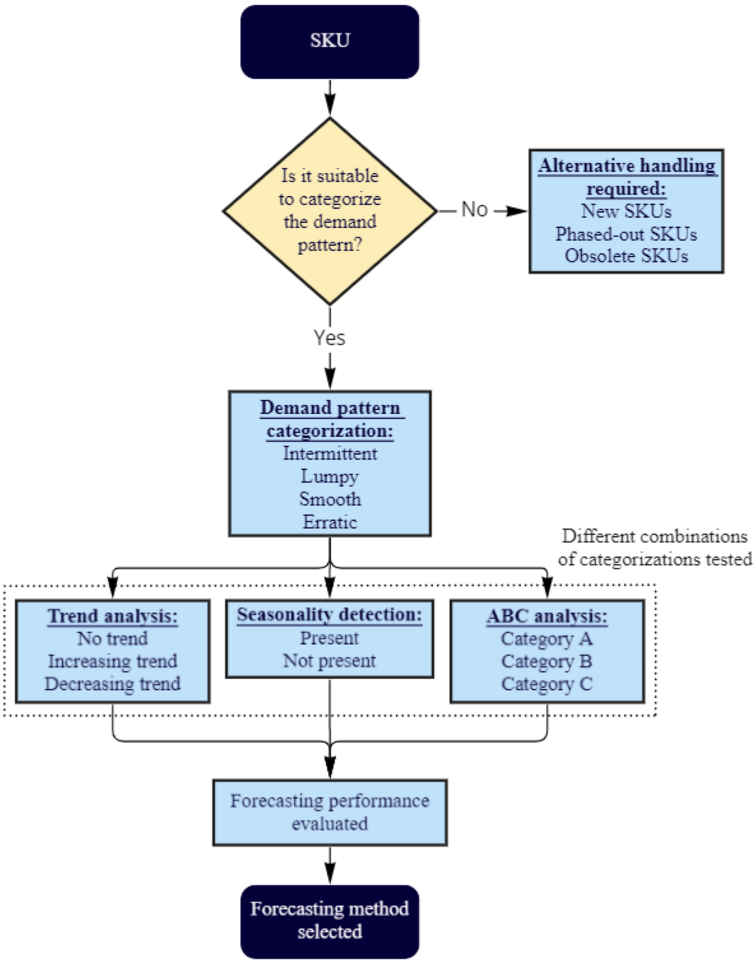

Figure 3.1: The proposed categorization process.

The first step of the proposed categorization process was to assess whether it was suitable or not to categorize the demand pattern of each SKU. New, phased-out and obsolete SKUs are inappropriate for the categorization process and were consequently not included in the research. Obsolete SKUs are inappropriate because they are not expected to have any future sales, which means that it is pointless to attempt to forecast their future demand. They were marked very clearly in the data and easily removed from further categorizations. Phased-out SKUs were harder to identify because the data from ÁTVR did not include inventory levels or orders placed. Decision-makers will in most cases know ahead of time which SKUs are phasing out and they can manually adjust the forecasting preferences according to the remaining inventory levels. This information was not available in the data and phased-out SKUs were consequently not included in the categorization process to avoid having to make assumptions about the initial inventory levels of each SKU. SKUs that did not have any demand for the last 12 months were considered phased-out and were therefore excluded from the analysis.

New SKUs are difficult to identify because it is unclear how long each SKU can realistically be considered new. Various rules-of-thumb have been proposed that use a fixed number of samples to categorize SKUs as new but they are misleading because they ignore the variability of the data (Hyndman & Athanasopoulos, 2021). The only limiting factors are the theoretical and practical minimum number of samples required for the categorizations and forecasting methods. The amount of data needed to determine which SKUs are new can then be adjusted to different circumstances and it becomes a managerial decision to quantify when there is enough data. SKUs in the analysis were considered new if there were less than 52 weeks since the first sale or if there were less than seven weeks with non-zero demand. This was done to ensure that the seasonal products that are only available for a few weeks per year were not categorized as new.

All SKUs that did not require alternative handling went through the subsequent steps of the categorization process. The ADI and CV^2 values were calculated for each SKU so that their demand pattern could be categorized according to the SBC categorization scheme. The traditional cut-off values of ADI = 1.32 and $CV^2 = 0.49$ were used to define the boundaries of the demand pattern quadrants.

The trend analysis was performed with the modified Mann-Kendall trend test from Hamed and Rao (1998). Their proposed modification addresses the autocorrelation issues with the original test and it has received significant attention in literature (Rutkowska, 2015). Other modifications, such as by Yue and Wang (2004), were considered because they proposed similar alterations, but the trend test from Hamed and Rao (1998) was ultimately used because it has a higher citation impact factor. A Python package from Hussain & Mahmud (2019) was then used to implement the modified Mann-Kendall trend test. The default parameters for the significance level and number of first significant lags were used in this case study, but they can of course be adjusted if necessary for other scenarios. The output from the test then categorized each SKU, based on the training data, as having no trend, increasing trend or decreasing trend.

The seasonality detection step of the categorization process was performed with the Kruskal-Wallis test. The data for each SKU was aggregated to a monthly level, but the requirement of at least five samples for the test to function properly meant that the monthly groups would have required five years of data. Instead, the months were combined to create four seasons: winter, spring, summer and autumn. Each season then contained three months of data per year, so a total of six samples were therefore used for each group since two years of data were included in the training set. The SciPy package in Python was used to implement the test in this case study. The default settings of the test were used and the output then indicated whether seasonality was detected for each SKU.

The ABC analysis was performed separately for each of the product categories at ÁTVR and it was based on the last 12 months of sales. Demand volume was then chosen as the criterion for two main reasons. The first reason for using demand volume was because ÁTVR already uses it as a criterion and the second reason was that the data for other criteria was either not collected or not as reliable. The prices in the data, for instance, only reflected the prices at the time the data was retrieved, so there was no way of knowing how often and how much the prices changed over the three-year period. The volume was split according to the traditional split of 80% for A, 15% for B and 5% for C products.

3.4.2 Forecasting process

The forecasting process involved finding appropriate forecasting methods and accuracy measures to test the performance of various combinations of categorizations. There were only several forecasting methods and accuracy measures implemented because it is practically impossible to implement all available forecasting methods (Spiliotis et al., 2020). The selection criteria for the statistical forecasting methods were that they were required to be well-established in literature and they had to have parameters that could be adjusted, otherwise the forecasting methods would always yield the same results for individual SKUs.

The CRO, SBA and TSB methods were selected as the statistical forecasting methods for the analysis. They are well-established in literature and have performed well in different situations for the SBC categorization scheme (Spiliotis et al., 2020). CRO and SBA were shown to display superior forecasting performance for the quadrants of the SBC categorization scheme shown in Figure 2.1. TSB is a more recent method that generally performs better than the other methods because it updates the estimates used in the calculations more frequently. It is therefore beneficial to include it in the analysis with CRO and SBA to compare how the different categorizations affect the forecasting performance. The methods were implemented in Python based on existing code that is publicly available, but the code was adjusted to the project.

RF was selected as the single ML method for this study. It was well-suited for the analysis because of its robustness and capability to accept categorical inputs. The results from the categorization process were used as categorical inputs for RF, along with the product categories and the volume of the containers. Features for last week's demand and the difference between the two most recent weekly demands were also created for a rolling four-week window to improve the accuracy of the forecasts. RF was implemented with the Scikit-learn package in Python (Pedregosa et al., 2012). Dummy encoding was performed for the categorical inputs and the default settings of the package were used to produce the results.

The forecasting methods were then used to forecast the weekly demand one step ahead for each SKU. The accuracy measures of MASE and RMSE were selected to evaluate the point forecast accuracy of the forecasting methods. They are both popular accuracy measures, despite having fundamental differences, and have been used in multiple studies to analyse the SBC categorization scheme (Bandeira et al., 2020; Doszyn, 2020; Heinecke et al., 2013). MASE is a scaled error that is generally applicable and has been shown to be well-suited for evaluating intermittent and lumpy demand (Hyndman & Koehler, 2006). RMSE, on the other hand, is a scale-dependent measure that is on the same scale as the data and estimates the average distance between predicted values and actual values.

The accuracy of each forecasting method was evaluated in several steps. The first step for the statistical methods involved testing a range of α and β parameter values in increments of 0.1 for a specific category. Predictions were then generated with one set of parameter values and the MASE and RMSE were calculated for each SKU within the category. Next, the average MASE and RMSE were calculated based on the values from all the individual SKUs in the category. Finally, the average MASE and RMSE for the category were compared to results generated by different α and β parameter values and only the best results were stored. For the ML method, all demand pattern quadrants were used as training data to build one model for each combination of categorizations. The demand patterns were all included in the same model because models are usually more powerful when they have larger training sets. Once the predictions had been made, the SKUs were separated into their respective demand pattern categories. Lastly, the average MASE and RMSE were calculated based on the individual SKUs from the same category.

4 Results

This chapter presents the results from the categorization process and forecasting process in separate subchapters.

4.1 Categorization process

The categorization process began once the data had been cleansed and prepared. An essential first step of the process was to categorize the underlying demand patterns of the SKUs according to the SBC categorization scheme. The ADI and CV^2 values were calculated for all SKUs in the data set and the traditional cut-off values were used to determine the demand pattern categories. Figure 4.1 shows a scatter plot of the demand pattern categorization, where each dot represents a SKU.

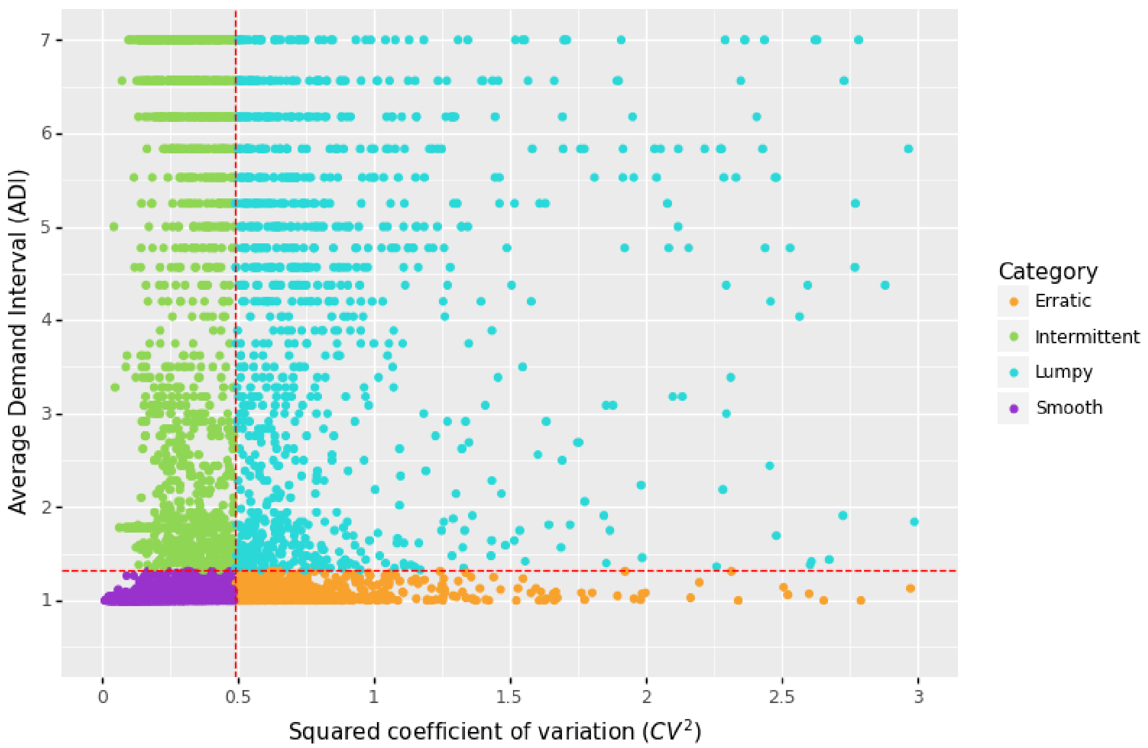

Figure 4.1. Scatter plot of the demand pattern categorization results.

The SKUs were largely concentrated around the intersection of the cut-off value boundaries. SKUs exhibit higher intermittency as the ADI values increase and higher erraticness as the CV^2 values increase. The ADI values of the SKUs in the data set were found to have higher variability than the CV^2 values. A higher proportion of ADI values therefore diverge from the average when compared to the CV^2 values. Additionally, a small number of SKUs were not shown in Figure 4.1 because the axis ranges were adjusted for the sake of clarity. Figure 4.2 then presents examples of weekly demand for typical SKUs in the data set to visualize the contrasts between the demand pattern categories.

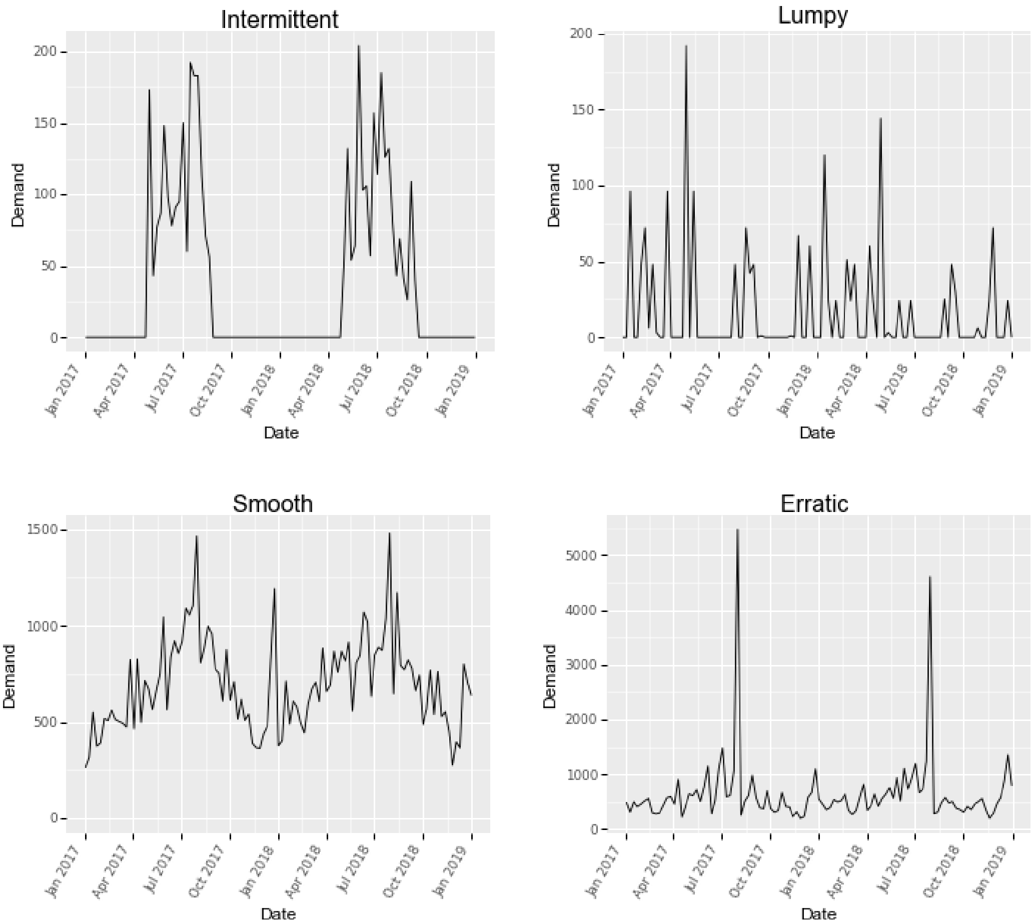

Figure 4.2. Examples of SKUs from each of the demand pattern quadrants in the SBC categorization scheme.

The intermittent demand pattern category contains SKUs with at least 26 weeks of zero demand that can occur at regular or irregular intervals throughout the two-year training period. Additionally, they have relatively stable demands for the periods of non-zero demand. SKUs in the lumpy category must also have at least 26 weeks of zero demand, but their demand fluctuates to a larger extent. The smooth category has the opposite features compared to the lumpy category. It therefore only contains SKUs with zero demand in fewer than 26 weeks and the non-zero demand periods have relatively stable demands. Similarly, the erratic category is the opposite of the intermittent category, so they have fewer than 26 weeks of zero demand and the demand fluctuates.

The visual distinction between demand patterns is often not as clear as Figure 4.2 suggests, but categorizations based on numerical values and cut-off boundaries remove all doubt. The demand pattern categorization results were summarized in Table 4.1, where the number of

SKUs in each category and the percentages of the total number of SKUs for each category were presented.

Table 4.1: Demand pattern categorization results.

Demand pattern	Number of SKUs	Percentage
Intermittent	1,514	24.6%
Lumpy	1,316	21.3%
Smooth	2,630	42.6%
Erratic	707	11.5%
Total	**6,167**	**100%**

The smooth demand pattern category had, by far, the highest number of SKUs in the data set. Contrarily, the erratic demand pattern category had the fewest number of SKUs. The intermittent and lumpy demand patterns had roughly half of the number of SKUs from the smooth category and around double the number from the erratic category. Approximately 45.9% of the SKUs had ADI values higher than 1.32, which means that they had at least 26 or more weeks of zero demand in the two-year training period. Furthermore, 32.8% of the SKUs had CV^2 values higher than 0.49.

The next step of the categorization process involved performing the trend analysis. The results from the modified Mann-Kendall trend test for each of the demand pattern categories can be seen in Table 4.2.

Table 4.2: Trend analysis results.

Demand pattern	No trend	Increasing trend	Decreasing trend
Intermittent	1,259	228	27
Lumpy	1,168	127	21
Smooth	1,582	421	627
Erratic	484	105	118
Total	**4,493**	**881**	**793**

The overwhelming majority of the SKUs did not have a statistically significant trend present in the training period. The smooth demand pattern category had the highest proportion SKUs with significant trend, either increasing or decreasing. It was closely followed by the erratic category, while the intermittent and lumpy categories had comparatively lower proportions. Roughly the same number of increasing and decreasing trends were identified for the erratic category. It was more common for the smooth category to have decreasing trends while the intermittent and lumpy categories had higher numbers of SKUs with increasing trends.

Next, the seasonality detection step was performed with the Kruskal-Wallis test. The preliminary results indicated that the test failed to identify several highly seasonal products, such as holiday specific products, that only had a few periods of high demand while the rest of the periods had zero demand. Additional code was therefore implemented to ensure the highly seasonal products were included in the seasonal category. All SKUs with eight or nine months of consecutive zero demand in both 2017 and 2018 were labelled as seasonal if they were not already. The seasonality detection results from the Kruskal-Wallis test and the additional check for consecutive zero demand periods can be seen in Table 4.3.

Table 4.3: Seasonality detection results.

Demand pattern	Present	Not present
Intermittent	1,155	359
Lumpy	991	325
Smooth	2,630	0
Erratic	707	0
Total	**5,483**	**684**

Significant seasonality was found to be present in all SKUs from the smooth and erratic demand pattern categories. The seasonality categorization was, as a result, not considered in the forecasting process for those demand patterns. The Kruskal-Wallis test did, however, not detect significant seasonality in 401 SKUs from the intermittent category and 461 from the lumpy category. The additional check for highly seasonal products later identified 42 intermittent SKUs and 136 lumpy SKUs that were moved to the category with seasonality present.

Finally, the ABC analysis was performed for the product categories in the data set. The ale category was the largest category before the data cleansing process, but the lager beer category contained nearly two-thirds of the SKUs after it. The mixed drinks and other beers categories had relatively few SKUs both before and after the data cleansing process. The ABC analysis results for the product categories were summarized in Table 4.4.

Table 4.4: ABC analysis results for the product categories.

Product category	Category A	Category B	Category C
Ale	219	418	668
Lager beer	572	1,071	2,384
Mixed drinks	84	139	169
Other beers	70	129	244
Total	**945**	**1,757**	**3,465**

The SKUs were divided into the ABC categories according to the traditional split for each product category and, in total, approximately 15.3% of the SKUs were categorized into category A, 28.5% into category B and 56.2% into category C. The resulting ABC categorizations were subsequently aggregated for the demand patterns and presented in Table 4.5.

Table 4.5: ABC analysis results for the demand patterns.

Demand pattern	Category A	Category B	Category C
Intermittent	105	273	1,136
Lumpy	55	165	1,096
Smooth	735	1,092	803
Erratic	50	227	430
Total	**945**	**1,757**	**3,465**

The smooth category had the lowest proportion of SKUs in category C and the highest proportion in category A. Furthermore, over three-quarters of the SKUs in category A had a smooth demand pattern while the other demand patterns had relatively few SKUs in category A. The overwhelming majority of the SKUs with an intermittent or lumpy demand pattern were in category C. The number of SKUs with an erratic demand pattern, on the other hand, did not spike or fluctuate to the same degree as the other demand patterns, and consistently increased from category A to category C.

4.2 Forecasting process

The forecasting process was divided into subchapters for each of the demand patterns from the SBC categorization scheme. The categorization process had already generated different groups of SKUs using established approaches and the forecasting process involved generating forecasts that were tailored to each group. The CRO, SBA, TSB and RF methods were used to compare the overall forecasting performance from specific categorizations to results from other categorizations with the same number of additional categorizations included. The top performing categorizations were subsequently highlighted for each distinct demand pattern.

As mentioned in Chapter 4.1, seasonality was found to be present in all SKUs with smooth or erratic demand patterns and the seasonality detection was consequently excluded from their forecasting process. The forecasts were generated for the demand patterns without any other adjustments or changes.

4.2.1 Intermittent demand

First, the forecasting results for SKUs that were categorized as having intermittent demand patterns were summarized in Table 4.6. The top performing single additional categorization was the trend analysis in terms of MASE and ABC analysis in terms of RMSE. The lowest MASE value was achieved with the TSB method and the lowest RMSE value with RF. When two additional categorizations were considered, the top performance was achieved with the trend analysis and seasonality detection in terms of MASE. Similarly, the trend analysis and ABC analysis led to the best performance in terms of RMSE. Again, the TSB method produced the lowest MASE value and RF the lowest RMSE value.

The TSB method produced consistently low MASE values compared to the other forecasting methods and was, in fact, the only method to achieve a MASE value with less than 1, meaning that it also outperformed the naïve approach. The RMSE values produced by the TSB method and RF were, however, quite similar but the lowest error produced by RF was slightly lower.

Table 4.6: Intermittent demand forecasting results.

Categorization	Forecasting method	Average MASE	Average RMSE
Demand pattern	CRO	2.709	100.186
	SBA	2.677	98.077
	TSB	0.973	67.279
	RF	1.523	71.942
Demand pattern Trend analysis	CRO	2.664	99.023
	SBA	2.628	94.881
	TSB	0.959	66.648
	RF	1.521	71.594
Demand pattern Seasonality	CRO	2.709	96.125
	SBA	2.608	97.353
	TSB	0.965	66.937
	RF	1.560	69.655
Demand pattern ABC analysis	CRO	2.661	96.654
	SBA	2.610	94.492
	TSB	0.973	66.905
	RF	1.148	66.100
Demand pattern Trend analysis Seasonality	CRO	2.663	95.063
	SBA	2.564	94.864
	TSB	0.955	66.542
	RF	1.558	69.595
Demand pattern Trend analysis ABC analysis	CRO	2.629	95.966
	SBA	2.560	92.794
	TSB	0.958	66.280
	RF	1.149	65.989
Demand pattern Seasonality ABC analysis	CRO	2.644	95.985
	SBA	2.562	93.880
	TSB	0.964	66.438
	RF	1.148	66.021
Demand pattern Trend analysis Seasonality ABC analysis	CRO	2.625	94.695
	SBA	2.517	91.740
	TSB	0.953	66.076
	RF	1.151	66.024

4.2.2 Lumpy demand

Next, the forecasting process was implemented for the lumpy demand pattern category and presented in Table 4.7.

Table 4.7: Lumpy demand forecasting results.

Categorization	Forecasting method	Average MASE	Average RMSE
Demand pattern	CRO	4.578	83.601
	SBA	3.496	80.634
	TSB	0.979	66.141
	RF	2.118	74.556
Demand pattern Trend analysis	CRO	4.556	83.094
	SBA	3.490	80.013
	TSB	0.971	65.638
	RF	2.085	74.269
Demand pattern Seasonality	CRO	4.051	83.601
	SBA	3.496	80.634
	TSB	0.979	65.768
	RF	2.073	70.426
Demand pattern ABC analysis	CRO	4.177	79.111
	SBA	3.496	80.237
	TSB	0.979	65.008
	RF	1.822	66.401
Demand pattern Trend analysis Seasonality	CRO	4.044	83.069
	SBA	3.489	79.988
	TSB	0.971	65.354
	RF	2.046	70.288
Demand pattern Trend analysis ABC analysis	CRO	4.155	79.000
	SBA	3.489	79.767
	TSB	0.971	64.799
	RF	1.826	65.988
Demand pattern Seasonality ABC analysis	CRO	3.813	78.977
	SBA	3.362	80.002
	TSB	0.979	64.827
	RF	1.881	66.715
Demand pattern Trend analysis Seasonality ABC analysis	CRO	3.804	78.652
	SBA	3.353	79.498
	TSB	0.970	64.610
	RF	1.868	66.623

The best single additional categorization was the trend analysis in terms of MASE and the ABC analysis in terms of RMSE, which was the same result as for the intermittent category.

However, the TSB method generated both the lowest MASE and RMSE values for the lumpy category. For two additional categorizations, the trend analysis and seasonality detection led to the same lowest MASE value as the trend analysis and ABC analysis. In a contrasting manner, the lowest RMSE value was reached unambiguously by the trend analysis and ABC analysis since they were the only combination of two additional categorizations to achieve it.

Both accuracy measures were noticeably higher for CRO and SBA, while TSB remained relatively similar when compared to the intermittent category results. The RMSE values for RF also remained relatively similar, but the MASE values increased for all categorizations.

4.2.3 Smooth demand

The results were then determined for the smooth demand pattern category and can be seen in Table 4.8.

Table 4.8: Smooth demand forecasting results.

Categorization	Forecasting method	Average MASE	Average RMSE
Demand pattern	CRO	0.954	151.425
	SBA	0.980	163.883
	TSB	0.851	145.721
	RF	0.993	172.182
Demand pattern Trend analysis	CRO	0.954	151.377
	SBA	0.980	163.805
	TSB	0.850	145.640
	RF	0.993	172.633
Demand pattern ABC analysis	CRO	0.952	151.197
	SBA	0.976	163.883
	TSB	0.848	145.528
	RF	0.960	173.503
Demand pattern Trend analysis ABC analysis	CRO	0.951	151.197
	SBA	0.975	163.652
	TSB	0.847	145.500
	RF	0.960	173.129

It was only possible to compare two different performances that were produced by a single additional categorization since the seasonality detection was excluded. The lowest MASE and RMSE values were achieved with the TSB method using the ABC analysis as the additional categorization. The TSB errors were considerably lower than the errors generated by the other methods. All forecasting methods from each of the categorizations did nevertheless outperform the naïve method with MASE values less than 1. Furthermore, the lowest RMSE values from the SKUs with a smooth demand pattern were significantly higher than the errors from the other demand patterns.

4.2.4 Erratic demand

Finally, the accuracy measures were calculated for all categorizations involving the erratic demand pattern category and the results can be seen in Table 4.9.

Table 4.9: Erratic demand forecasting results.

Categorization	Forecasting method	Average MASE	Average RMSE
Demand pattern	CRO	0.996	55.978
	SBA	0.942	55.882
	TSB	0.853	54.718
	RF	1.161	65.734
Demand pattern Trend analysis	CRO	0.996	55.825
	SBA	0.939	55.843
	TSB	0.852	54.684
	RF	1.176	66.206
Demand pattern ABC analysis	CRO	0.993	55.945
	SBA	0.932	55.852
	TSB	0.851	54.682
	RF	1.058	65.270
Demand pattern Trend analysis ABC analysis	CRO	0.990	55.783
	SBA	0.928	55.758
	TSB	0.850	54.535
	RF	1.050	64.921

As with the smooth category, the seasonality detection was not included as one of the categorizations for the erratic category. The forecasting performance comparison was therefore between the trend analysis and the ABC analysis. The top performances for each categorization were nearly identical for both accuracy measures, but the ABC analysis produced slightly lower MASE and RMSE values. Both lowest errors were achieved with the TSB method and RF was the only method to not outperform the naïve method since the MASE values were always greater than 1.

5 Discussion

There are countless ways to categorize products and to tune forecasting parameters for different forecasting methods. The systematic categorization process proposed in this study was based on several established approaches and grouped SKUs with similar characteristics together. Various combinations of categorizations then generated different groups of SKUs and the forecasting performance of several forecasting methods was analysed for each group by testing a range of forecasting parameters. The top performing categorizations were highlighted in Chapter 4.2 for each underlying demand pattern, to facilitate the selection of appropriate forecasting methods.

5.1 Categorization process

The categorization process was initiated by determining the underlying demand pattern of SKUs with the SBC categorization scheme. The resulting number of SKUs in each category indicated that just over two-thirds of the SKUs had relatively stable demands with low CV^2 values. Furthermore, roughly half of the SKUs were above the ADI cut-off value while the other half were below it. The data cleansing process influenced the results because it was performed with specific assumptions that limited the size the data set. The assumptions were relevant to the data set from ÁTVR and ensured all SKUs had the same information available, but the downside was that the cleansed data set might not have provided an accurate representation of the full data set. For instance, the number of SKUs with an intermittent or lumpy demand pattern would have been significantly higher if SKUs with missing feature information were included in the cleansed data set. Certain feature information can be inapplicable for some SKUs and is for that reason often not recorded. The SBC categorization scheme only requires time series data to function and is therefore unaffected by missing feature information. However, the missing information could have negatively affected the RF method later in the forecasting process and was therefore excluded from the analysis for the sake of simplicity. It is consequently important to consider the implications of the assumptions made in the data cleansing process and to define precisely what information is necessary before interpreting the results.

Additional approaches were then used to further categorize the SKUs in the data set, starting with the trend analysis. The results showed that most of the SKUs did not have a significant trend present in the training data. This indicated that the demand for most of the alcoholic drinks at ÁTVR did not change drastically over the two-year training period. This interpretation is, however, tied to the selected significance level of the trend test, because it could be adjusted to produce different results. Furthermore, the results were also dependent on the number of non-zero samples for each SKU in the training set. It was impossible to distinguish between stock-outs and periods of zero demand, so only the non-zero samples could indicate whether significant changes occurred over time.

Seasonality was detected for the vast majority of the SKUs in the data set. It is reasonable to assume that most SKUs in the data set had seasonal demands because the demand for alcoholic drinks generally fluctuates between seasons. The seasonality detection step did, however, fail to identify highly seasonal SKUs with the Kruskal-Wallis test. The test was found to have low power when the input for the seasons contained multiple zeros. This was likely caused by the high number of ties in the data, so the rank-based test was unable to assign unique rank numbers. As a result, an additional seasonality check was implemented to ensure the highly seasonal SKUs were placed in the proper category. An alternative approach that measures the strength of seasonality could likely have produced more evenly sized groups of SKUs, which would have been better suited for this data set.

The ABC analysis for the product categories of the data set showed that most of the SKUs in the A category had a smooth demand pattern. This makes sense because the A category only contains SKUs with the highest cumulative demand for the past year and SKUs in the smooth category have a stable demand that occurs at regular intervals. SKUs with consistently high demand are therefore both in the A category and have a smooth demand pattern. Not surprisingly, most of the SKUs with an intermittent or lumpy demand pattern were in the C category. These SKUs have relatively sparse sales histories and are unlikely to be sold in large quantities for an extended period.

5.2 Forecasting process

Various groups of SKUs had been created in the categorization process and the selected forecasting methods were used to predict their future demand. The forecasting performances achieved by grouping SKUs according to the different categorizations were only compared to other results with the same number of additional categorizations. The purpose of this was to allow for a fair comparison between the different categorizations since the forecasting performance of the statistical methods continuously improves as the groups of SKUs become smaller. The results produced by the statistical methods could not get any worse by splitting a group into two smaller groups because the same parameters could always be chosen if better parameters were not found. The parameters for the selected forecasting methods were by no means perfectly optimized, but the range of parameters that were tested enabled the methods to improve the forecasting performance. ML methods, on the other hand, do not necessarily benefit from generating forecasts for smaller groups of SKUs, which the RF results indicated in this study.

A larger number of groups with fewer SKUs in each one produces better results for the statistical methods, but also increases the time needed to generate predictions for the whole data set. The time needed to test all possible forecasting methods and parameters was, however, not a good indicator of performance because the test would only need to be performed once or very infrequently. Only a small subset of the top performing forecasting methods and parameters would be tested in practice for each group, so measuring the time for all combinations would not have produced relevant results. Run time was consequently not measured in this study since it would not have represented the actual time it took to generate the results from the appropriate subset.

The top performing categorizations were identified and highlighted for each of the distinct demand patterns. The trend analysis was shown to produce the lowest MASE value for the intermittent and lumpy demand patterns when one additional categorization was included. The lowest MASE value for the smooth and erratic demand patterns were, however, generated with the ABC analysis. Moreover, the lowest RMSE value for each of the demand patterns was achieved with the ABC analysis. The results for two additional categorizations showed that the combination of trend analysis and seasonality detection produced the lowest MASE value for intermittent and lumpy demand. Additionally, the combination of trend analysis and ABC analysis generated the same lowest MASE value for lumpy demand and it also produced the lowest RMSE value for both intermittent and lumpy demand.

The forecasting performance of the additional categorizations with seasonality detection was generally worse than the other approaches. This could likely be contributed to the disparity between the sizes of the subcategories. Seasonality was detected for most of the SKUs in the data set while only a small number of SKUs did not have seasonality present. As a result, a larger proportion of the SKUs was categorized to a single subcategory compared to the results from the other categorizations.

The results from the CRO and SBA forecasting methods support the findings from previous studies. CRO produced better results for the smooth demand category while SBA outperformed CRO for the other demand patterns. The differences between the performances of the forecasting methods were less apparent for SKUs with low ADI values. They had relatively few periods of zero demand and therefore mostly avoided the disadvantage of CRO and SBA, which is that they only update their forecasts in periods of non-zero demand. This means that they continue to produce the same forecast until the next period of non-zero demand. The resulting weekly forecasting errors could therefore be consistently high if the last forecast before multiple periods of zero demand is high. TSB, on the other hand, updates the forecasts in every period and avoids this issue. This was likely the reason why the average MASE for the intermittent and lumpy demand pattern categories was considerably worse using CRO and SBA compared to TSB. The naïve prediction from MASE quickly adjusts to periods of zero demand, since it always bases the next prediction on the demand from the previous period, and TSB was the only method to consistently outperform it for all demand patterns.

The RMSE accuracy measure amplifies higher errors more than lower errors and, as a result, penalizes more for large errors compared to the in-sample MAE in MASE. The contrast between the accuracy measures is essentially that MASE measures performance relative to the naïve method while RMSE measures the magnitude of the error. RF had higher magnitudes of errors on average compared to the statistical methods for the smooth and erratic demand patterns. A possible explanation for the weaker performance of RF compared to the statistical methods for those demand patterns is that RF cannot extrapolate. Predictions for regression tasks are generated by the average value of multiple independent predictions, which means that predictions are limited by the highest and lowest samples in the training set. If the test set contains periods of higher or lower demand than the training set, then RF is unable to generate a precise prediction.

6 Conclusion

The SKUs in the data set were categorized according to the SBC categorization scheme, where 1,514 SKUs were determined to have an intermittent demand pattern, 1,316 were lumpy, 2,630 were smooth and 707 were erratic. A trend analysis with the modified Mann-Kendall trend test, seasonality detection with the Kruskal-Wallis test and an ABC analysis were then performed to further categorize the SKUs. The additional categorizations produced various groups of SKUs with similar characteristics and the future demand of each group was forecasted to determine which categorization led to the lowest demand forecasting error for each demand pattern. Due to their prominence in literature, CRO, SBA, TSB and RF were then selected as the forecasting methods for this study.

The top performing categorizations were identified for each of the distinct demand patterns by testing a range of parameters for the selected forecasting methods and comparing the results based on the accuracy measures of MASE and RMSE. For a single additional categorization, the trend analysis led to the lowest MASE value while the ABC analysis led to the lowest RMSE value for the intermittent and lumpy demand patterns. Furthermore, the ABC analysis yielded both the lowest MASE and RMSE value for the smooth and erratic demand patterns. For two additional categorizations, the trend analysis and seasonality detection produced the lowest MASE value for the intermittent and lumpy demand patterns. The combination of the trend analysis and ABC analysis achieved an equally low MASE value for the lumpy demand pattern, but also produced the lowest RMSE value for both the intermittent and lumpy demand patterns. The seasonality detection step was excluded from the analysis for the smooth and erratic demand patterns and comparisons between different combinations of two additional categorizations were therefore not possible for those demand patterns.

The proposed systematic categorization process established a foundation of approaches that facilitate the selection of forecasting methods for multiple clearly defined groups of SKUs. The different categories each SKU belonged to could be used to quickly identify a subset of forecasting methods and parameters likely to produce the lowest forecasting error. A new SKU with adequate data could therefore be categorized and assigned to a relevant group according to the categorization process, where only a small subset of the top performing forecasting methods and parameters would be tested for that specific group. The interpretability of the categorization process was evidently very high since simply checking the resulting categories of each SKU sufficed to clarify how an appropriate forecasting method was selected.

Numerous factors can influence forecasting performance and it is often challenging to accurately forecast the future demand of multiple SKUs. Moreover, it is imperative to maintain a good balance between acceptable forecasting performance and the time needed to generate forecasts to be able to quickly adjust in an ever-changing world. The systematic categorization process proposed in this study serves as the first step towards finding an advantageous balance. Categorization methods function in diverse ways and can support demand forecasting tasks when a suitable combination of categories is identified for a given set of circumstances.

6.1 Limitations

A significant limitation of the study was that only a small number of additional categorizations and forecasting methods were tested. A more extensive analysis is needed to expand the categorization process and to verify the obtained results. Furthermore, the generalizability of the results could not have been thoroughly estimated since only a single data set was used in the study. The analysis should consequently be repeated for different data sets to investigate the applicability of the presented categorization process.

The weekly forecasts were only generated one step ahead and the results could therefore not indicate whether there was a significant difference in forecasting performance for short-term and slightly longer forecasting horizons. SKUs are, in practice, rarely only forecasted one week in advance and it becomes more important to estimate demand further in the future when inventory management is considered. Moreover, it was impossible to distinguish between periods of real stock-outs and zero demand periods because information about inventory levels was not available. As a result, the categorizations for some SKUs were possibly misleading since there is a fundamental difference between not having a potentially highly sought-after SKU in stock and simply having no demand for a SKU that was available in stock.

6.2 Future research

There are many interesting options for future research that warrant further investigation. One such option would be to extend the categorization process to include new products by utilizing ML techniques. A cluster analysis could then be performed, where products are grouped together based on the similarities of their features. The results would be used to estimate the most appropriate categories from the categorization process for all new products. The future demand could then be forecasted with the support of the estimated categorizations, or with additional steps to enhance the forecasting process, until products finally reach the minimum sample size required to go through the categorization process.

Another noteworthy possibility involves measuring the time it takes to generate the forecasts for each group of SKUs. The time measurements would only be based on a small subset of forecasting methods and parameters that had been determined to be suitable for the group. Researchers could then investigate the trade-off between forecasting performance and the run time needed to produce forecasts. On the one hand, the parameter tuning for small groups of SKUs can be customized to a greater extent than for larger groups, which likely improves forecasting performance. On the other hand, the run time increases as groups become smaller since a wider array of forecasting methods and parameters need to be tested for the whole data set.

Future research could also further develop the categorization process by including more categorizations and comparing the relative importance of each categorization. The results could then be used to define guidelines for selecting a suitable number of categorizations for different data sets. Furthermore, it would allow decision-makers to highlight certain groups of SKUs with combinations of traits that could be worthwhile to monitor more closely. Inventory implications would also be important to consider and the most cost-effective inventory management decisions could be achieved with a more relevant performance indicator. Another possibility would be to emphasize enhancing the forecasting performance by increasing the number of forecasting methods in the analysis, both for statistical and ML methods, and by testing more sophisticated methods. The most appropriate methods could then be determined for each category and only a specific group of them would be used in practice to reduce the run time of the forecasting process.

References

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. https://doi.org/10.1016/j.ijpe.2018.01.026

Bacchetti, A., & Saccani, N. (2012). Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice. *Omega*, *40*(6), 722–737. https://doi.org/10.1016/j.omega.2011.06.008

Bandeira, S. G., Alcalá, S. G. S., Vita, R. O., & Barbosa, T. M. G. de A. (2020). Comparison of selection and combination strategies for demand forecasting methods. *Production*, *30*, Article e20200009. https://doi.org/10.1590/0103-6513.20200009

Boylan, J. E., & Syntetos, A. A. (2021). *Intermittent demand forecasting: Context, methods and applications*. Wiley.

Boylan, J. E., Syntetos, A. A., & Karakostas, G. C. (2008). Classification for forecasting and stock control: A case study. *Journal of the Operational Research Society*, *59*(4), 473–481. https://doi.org/10.1057/palgrave.jors.2602312

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brillinger, D. R. (1994). Trend analysis: Time series and point process problems. *Environmetrics*, *5*(1), 1–19. https://doi.org/10.1002/env.3170050102

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time Series and Forecasting* (3rd ed.). Springer.

Chase, C. (2021). *Consumption-based forecasting and planning: Predicting changing demand patterns in the new digital economy*. Wiley.

Ching-Chin, C., Ka Ieng, A. I., Ling-Ling, W., & Ling-Chieh, K. (2010). Designing a decision-support system for new product sales forecasting. *Expert Systems with Applications*, *37*(2), 1654–1665. https://doi.org/10.1016/j.eswa.2009.06.087

Costantino, F., Di Gravio, G., Patriarca, R., & Petrella, L. (2018). Spare parts management for irregular demand items. *Omega*, *81*, 57–66. https://doi.org/10.1016/j.omega.2017.09.009

Croston, J. D. (1972). Forecasting and Stock Control for Intermittent Demands. *Operational Research Quarterly (1970-1977)*, *23*(3), 289–303. https://doi.org/10.2307/3007885

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*(3), 510–522. https://doi.org/10.1016/j.ijforecast.2012.09.002

de Oliveira, A. C. A., Jorge, J. M., dos Santos, A. C., & Filho, G. P. R. (2020). Neural Network with Specialized Knowledge for Forecasting Intermittent Demand. *Transdisciplinary Engineering for Complex Socio-Technical Systems – Real-Life Applications*, *12*, 524–533. https://doi.org/10.3233/ATDE200113

Doszyn, M. (2020). Accuracy of Intermittent Demand Forecasting Systems in the Enterprise. *European Research Studies Journal*, *23*(4), 912–930. https://doi.org/10.35808/ersj/1723

Douissa, M. R., & Jabeur, K. (2016). A New Model for Multi-criteria ABC Inventory Classification: PROAFTN Method. *Procedia Computer Science*, *96*, 550–559. https://doi.org/10.1016/j.procs.2016.08.233

Eaves, A. H. C., & Kingsman, B. G. (2004). Forecasting for the ordering and stock-holding of spare parts. *Journal of the Operational Research Society*, *55*(4), 431–437. https://doi.org/10.1057/palgrave.jors.2601697

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society*, *59*(9), 1150–1172. https://doi.org/10.1057/palgrave.jors.2602597

Gauthier, T. D., & Hawley, M. E. (2015). Statistical Methods. In B. L. Murphy & R. D. Morrison (Eds.), *Introduction to Environmental Forensics* (3rd ed., pp. 99–148). Elsevier. https://doi.org/10.1016/B978-0-12-404696-2.00005-9

Gelei, A., & Dobos, I. (2020). Forecasting of Sporadic Products: Practical and Theoretical Considerations. *Periodica Polytechnica Social and Management Sciences*, *28*(2), 101–110. https://doi.org/10.3311/PPso.13170

Gutierrez, R. S., Solis, A. O., & Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. *International Journal of Production Economics*, *111*(2), 409–420. https://doi.org/10.1016/j.ijpe.2007.01.007

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1–4), 182–196. https://doi.org/10.1016/S0022-1694(97)00125-X

Harvey, A. (2016). Trend Analysis. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (pp. 1–21). Wiley. https://doi.org/10.1002/9781118445112.stat07817.pub2

Heinecke, G., Syntetos, A. A., & Wang, W. (2013). Forecasting-based SKU classification. *International Journal of Production Economics*, *143*(2), 455–462. https://doi.org/10.1016/j.ijpe.2011.11.020

Hirsch, R. M., & Slack, J. R. (1984). A Nonparametric Trend Test for Seasonal Data With Serial Dependence. *Water Resources Research*, *20*(6), 727–732. https://doi.org/10.1029/WR020i006p00727

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, *18*(1), 107–121. https://doi.org/10.1029/WR018i001p00107

Hussain, Md., & Mahmud, I. (2019). pyMannKendall: A python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software*, *4*(39), 1556. https://doi.org/10.21105/joss.01556

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting*, *36*(1), 167–177. https://doi.org/10.1016/j.ijforecast.2019.05.008

Johnston, F. R., & Boylan, J. E. (1996). Forecasting for Items with Intermittent Demand. *The Journal of the Operational Research Society*, *47*(1), 113–121. https://doi.org/10.2307/2584256

Kendall, M. G. (1955). *Rank Correlation Methods* (2nd ed.). Charles Griffin.

Kharfan, M., Chan, V. W. K., & Efendigil, T. F. (2021). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. *Annals of Operations Research*, *303*(1–2), 159–174. https://doi.org/10.1007/s10479-020-03666-w

Kiefer, D., Grimm, F., Bauer, M., & Van, D. (2021). Demand Forecasting Intermittent and Lumpy Time Series: Comparing Statistical, Machine Learning and Deep Learning Methods. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1425–1434. https://doi.org/10.24251/HICSS.2021.172

Kostenko, A. V., & Hyndman, R. J. (2006). A note on the categorization of demand patterns. *Journal of the Operational Research Society*, *57*(10), 1256–1257. https://doi.org/10.1057/palgrave.jors.2602211

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, *73*(5), 937–954. https://doi.org/10.1080/01605682.2021.1892464

Krajewski, L. J., Malhotra, M. K., & Ritzman, L. P. (2019). *Operations management. Processes and supply chains* (12th ed.). Pearson.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Liu, P. (2020). Intermittent demand forecasting for medical consumables with short life cycle using a dynamic neural network during the COVID-19 epidemic. *Health Informatics Journal*, *26*(4), 3106–3122. https://doi.org/10.1177/1460458220954730

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. https://doi.org/10.1016/j.ijpe.2016.10.021

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), Article e0194889. https://doi.org/10.1371/journal.pone.0194889

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, *13*(3), 245–259. https://doi.org/10.2307/1907187

Nikolopoulos, K. (2021). We need to talk about intermittent demand forecasting. *European Journal of Operational Research*, *291*(2), 549–559. https://doi.org/10.1016/j.ejor.2019.12.046

Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and Application of the Kruskal-Wallis Test. *Applied Mechanics and Materials*, *611*, 115–120. https://doi.org/10.4028/www.scientific.net/AMM.611.115

Patterson, K. D. (2012). *Unit Roots Tests in Time Series Volume 2: Extensions and Developments*. Palgrave Macmillan.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). *Scikit-learn: Machine Learning in Python*. https://doi.org/10.48550/ARXIV.1201.0490

Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, *66*(6), 914–924. https://doi.org/10.1057/jors.2014.62

Placencia, I. A., Sánchez-Partida, D., Martínez-Flores, J. L., & Cano-Olivos, P. (2021). Sales forecast for aggregate planning: Case study of an industrial products company in Mexico. *Acta Logistica*, *8*(4), 381–392. https://doi.org/10.22306/al.v8i4.240

Ravinder, H., & Misra, R. B. (2014). ABC Analysis For Inventory Management: Bridging The Gap Between Research And Classroom. *American Journal of Business Education (AJBE)*, *7*(3), 257–264. https://doi.org/10.19030/ajbe.v7i3.8635

Regattieri, A., Gamberi, M., Gamberini, R., & Manzini, R. (2005). Managing lumpy demand for aircraft spare parts. *Journal of Air Transport Management*, *11*(6), 426–431. https://doi.org/10.1016/j.jairtraman.2005.06.003

Rezaei, J., & Dowlatshahi, S. (2010). A rule-based multi-criteria approach to inventory classification. *International Journal of Production Research*, *48*(23), 7107–7126. https://doi.org/10.1080/00207540903348361

Rožanec, J. M., Kažič, B., Škrjanc, M., Fortuna, B., & Mladenić, D. (2021). Automotive OEM Demand Forecasting: A Comparative Study of Forecasting Algorithms and Strategies. *Applied Sciences*, *11*(15), 6787. https://doi.org/10.3390/app11156787

Rutkowska, A. (2015). Properties of the Cox–Stuart Test for Trend in Application to Hydrological Series: The Simulation Study. *Communications in Statistics - Simulation and Computation*, *44*(3), 565–579. https://doi.org/10.1080/03610918.2013.784988

Shcherbakov, M., Brebels, A., Shcherbakova, N. L., Tyukov, A., Janovsky, T. A., & Kamaev, V. A. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, *24*(24), 171–176. https://doi.org/10.5829/idosi.wasj.2013.24.itmies.80032

Spiliotis, E., Makridakis, S., Kaltsounis, A., & Assimakopoulos, V. (2021). Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data. *International Journal of Production Economics*, *240*, Article 108237. https://doi.org/10.1016/j.ijpe.2021.108237

Spiliotis, E., Makridakis, S., Semenoglou, A.-A., & Assimakopoulos, V. (2020). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research*, 1–25. https://doi.org/10.1007/s12351-020-00605-2

Syntetos, A. A., Babai, M. Z., Lengu, D., & Altay, N. (2011). Distributional Assumptions for Parametric Forecasting of Intermittent Demand. In N. Altay & L. A. Litteral (Eds.), *Service Parts Management* (pp. 31–52). Springer London. https://doi.org/10.1007/978-0-85729-039-7_2

Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics*, *71*(1–3), 457–466. https://doi.org/10.1016/S0925-5273(00)00143-2

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*(2), 303–314. https://doi.org/10.1016/j.ijforecast.2004.10.001

Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society*, *56*(5), 495–503. https://doi.org/10.1057/palgrave.jors.2601841

Teunter, R. H., Babai, M. Z., & Syntetos, A. A. (2010). ABC Classification: Service Levels and Inventory Costs: ABC Classification. *Production and Operations Management*, *19*(3), 343–352. https://doi.org/10.1111/j.1937-5956.2009.01098.x

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. https://doi.org/10.1016/j.ejor.2011.05.018

Thomopoulos, N. T. (2015). *Demand Forecasting for Inventory Control*. Springer. https://doi.org/10.1007/978-3-319-11976-2

Tsao, Y.-C., Kurniati, N., Pujawan, I. N., & Yaqin, A. M. (2019). Spare Parts Demand Forecasting in Energy Industry: A Stacked Generalization-Based Approach. *Proceedings of the 2019 International Conference on Management Science and Industrial Engineering - MSIE 2019*, 163–167. https://doi.org/10.1145/3335550.3335573

Turrini, L., & Meissner, J. (2019). Spare parts inventory management: New evidence from distribution fitting. *European Journal of Operational Research*, *273*(1), 118–130. https://doi.org/10.1016/j.ejor.2017.09.039

Vandeput, N. (2021). *Data science for supply chain forecasting* (2nd ed.). De Gruyter. https://doi.org/10.1515/9783110671124

Varghese, V., & Rossetti, M. (2008). A Classification Approach for Selecting Forecasting Techniques for Intermittent Demand. *Proceedings of the 2008 Industrial Engineering Research Conference*, 863–868.

Wallström, P., & Segerstedt, A. (2010). Evaluation of forecasting error measurements and techniques for intermittent demand. *International Journal of Production Economics*, *128*(2), 625–636. https://doi.org/10.1016/j.ijpe.2010.07.013

Williams, T. M. (1984). Stock Control with Sporadic and Slow-Moving Demand. *Journal of the Operational Research Society*, *35*(10), 939–948. https://doi.org/10.1057/jors.1984.185

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE ’14*, 1–10. https://doi.org/10.1145/2601248.2601268

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.

Yue, S., & Wang, C. (2004). The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series. *Water Resources Management*, *18*(3), 201–218. https://doi.org/10.1023/B:WARM.0000043140.61082.60