Short lifecycle items detection for demand forecasting

João Pedro Neves

Dissertação de Mestrado

Orientador na FEUP: Prof. Gonçalo Figueira

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado em Engenharia e Gestão Industrial

2023-06-26

"Prediction is very difficult, especially if it's about the future."

Nils Bohr

Resumo

Otimização é um processo que tem crescido em relevância para várias indústrias ao longo das últimas décadas. As empresas devem procurar a melhoria continua das suas operações para manter a relevância e prosperar. Os mercados, no retalho, são marcados por forte competitividade e padrões de consumidor em rápida transição.

Produtos de ciclo de vida curto são uma classe emergente que demonstra janelas de relevância cada vez menores. Estes são tradicionalmente associados a *fashion*, mas tem-se tornado prevalentes noutros setores, nomeadamente produtos tecnológicos, entretenimento ou até alguns bens essenciais. A previsão de procura tem um papel fundamental perante incertezas e no planeamento eficiente de operações de *replenishment* e *supply chain*. A precisão de *forecast* é essencial para a otimização de operações de negócio ou antecipar mudanças de tendência. No entanto a previsão enfrenta um obstáculo quando perante produtos de ciclo de vida curto, caracterizados por elevada variação na procura e insuficiência de dados.

Apesar de diversos modelos baseados em métodos estatísticos ou algoritmos de *machine learning* terem sido propostos para adaptar a previsão de procura às características dos produtos de ciclo de vida curto, a proposta de procedimentos para identificar esta classe de produtos na literatura é limitada. No entanto, esta classificação é essencial para adequadamente selecionar o modelo e parâmetros de previsão ideais e assim melhorar a performance de previsão.

A presente dissertação é desenvolvida no contexto de uma empresa de consultoria de IT (*information technology*) que apoia os retalhistas na implementação e configuração de software orientado ao retalho. Nos projetos implementados na Retail Consult, as soluções de previsão de procura baseiam-se na ferramenta Oracle *Retail Demand Forecasting* (RDF). Esta solução conta com métodos estatísticos e de *machine learning* que visam a previsão de procura com exatidão num ambiente integrado. As funcionalidades do RDF permitem a execução de procedimentos especificamente otimizados para prever a procura de produtos de ciclo de vida curto. No entanto, estes devem ser determinados pelo utilizador. Assim, um procedimento estruturado e automático para analisar e classificar produtos como de ciclo de vida curto ou longo pode melhorar significativamente a eficiência e precisão deste processo, com o objetivo de obter previsões de confiança em tempo útil. O modelo proposto aplica-se à melhoria do processo de previsão, no entanto traz avanços para a classificação e análise de produtos de ciclo de vida curto.

A presente dissertação propõe um novo modelo de classificação com procedimentos estruturados para identificar produtos de ciclo de vida curto e separá-los dos de ciclo de vida longo. Com isso em mente, os métodos de *clustering*, *k-means* e *DBSCAN*, adaptados a séries temporais, são propostos e implementados de acordo com procedimentos de preparação de dados com o objetivo de agrupar produtos de acordo com a semelhança dos comportamentos da procura. A metodologia apresentada resultou na formação de dois *clusters* classificados como de ciclo de vida curto ou longo. Após uma extensa avaliação de métricas de previsão e indicadores (coeficiente de variação, intermitência, entre outros), numa amostra de 1000 *stock keeping units*.

A adoção do modelo proposto é motivada pelas conclusões obtidas pela metodologia de avaliação. Os resultados sugerem forte correlação entre a classificação obtida e os indicadores de procura e métricas de *forecast* aceitáveis no setor. Isto sugere que as classes SLC e LLC geradas representam os produtos de ciclo de vida curto e longo, respetivamente. Á medida que os produtos de ciclo de vida curto se tornam potenciadores para a maioria das indústrias, a adoção de um procedimento estruturado, como o proposto no presente trabalho, é crucial para otimizar as atividades operacionais e de marketing.

Short lifecycle items detection for demand forecasting

Abstract

Optimization is increasing in relevance across the tasks of companies over the last decades. Companies must seek to continuously improve operations to remain relevant and prosper. The retail markets are fiercely competitive and consumer patterns shift quickly, now more than ever.

Short life cycle items are an emerging class of products across different sectors. Traditionally associated with fashion, SLC (short life cycle) items have become prevalent in other industries such as technology, entertainment, or even food products. Due to the lack of historical data and extreme sales variability in most SLC, forecasting the demand of these products is a complex task. However, demand forecasting has a crucial role when dealing with uncertainties and efficiently manage replenishment and supply chain operations. Accurate demand forecasts are key to optimize business tasks or anticipate trend shifts.

Even though diverse methods have been proposed to adequately forecast SLC items demand based on statistical or machine learning algorithms, the process of identifying SLC items within the assortment is not referenced in the literature. However, this classification is of pivotal importance to adequately select the best model and parameters for forecasting.

This thesis is developed in the context of an IT (information technology) consulting company that supports retailers in the implementation and configuration of software oriented to retail. In Retail Consult, forecasting solutions are based on the RDF tool. This solution relies on multiple statistical methods and machine learning algorithms to accurately forecast demand in an integrated environment. RDF features procedures specifically tailored to handle SLC demand, however, these labels must be provided by the user. A structured automated approach to classify and analyse assortments as SLC and LLC (long life cycle) classes can significantly improve the efficiency and accuracy of this process, and bottom line achieve more reliable and timely forecasts. The proposed model is valuable for forecasting purposes but also to obtain an insightful new classification of the products.

The present dissertation proposes a novel classification framework to identify SLC items and separate them from LLC. For that matter, clustering algorithms adapted for time series data, k-means and DBSCAN, are implemented along with pre-processing techniques and cluster classification aiming to group products according to their demand behaviours. This methodology results in the formation of two clusters that are assigned as SLC and LLC, after an extensive review of forecast metrics and indicators over a sample of 1000 stock keeping units.

The adoption of the proposed framework is motivated by the positive conclusions drawn from the evaluation methodology. The results are highly correlated with the linear combination with thresholds of the demand indicators and forecast metrics. This indicates that the SLC and LLC classes adequately represent the actual SLC and LLC items. As SLC items become key drivers for many industries, the adoption of a structured procedure to identify them, as the hereby proposed, is crucial for streamlining operational and marketing activities.

Keywords: short life cycle, demand forecasting, partition, and density-based clustering

Acknowledgments

Obrigado à Retail Consult pela oportunidade de contribuir para um projeto ambicioso e de grande aprendizagem, sempre rodeado de uma equipa divertida e muito apoio!

Obrigado ao Gonçalo Figueira por guiar o trabalho e o apoiar nas decisões críticas ao longo das várias fases do projeto.

Obrigado ao Ruben Pais pela boa disposição e forte gestão do projeto. Uma pessoa chave para o seu sucesso e sempre disponível para ajudar e partilhar ideias.

Obrigado ao Carlos Lima, Paulo Marques e Ana Rita Novo por contribuírem para a criação do projeto, e pela presença ao longo do trabalho.

Um especial agradecimento aos meus pais e à minha irmã que, não só me apoiaram e acompanharam ao longo destes 5 anos, como também nos que vieram antes.

À Maria pelo carinho e apoio incondicional, e por me rodear de amigos caninos durante o projeto.

A todos os meus bons amigos de faculdade e de outros meios que sempre aparecem e criam bons momentos.

João Neves

Contents

Abbreviations

PLC	Product Life Cycle
SLC	Short Life Cycle
LLC	Long Life Cycle
SKU	Stock keeping Unit
RDF	(Oracle) Retail Demand Forecasting.
SES	Simple Exponential Smoothing
HW	Holt-Winters
ARIMA	Auto Regressive Integrated Moving Averages
CV	Coefficient of Variation
MSE	Mean Square Error
ME	Mean Error
MAE	Mean Absolute Error
SMAPE	Symmetric Mean Absolute Error
DBSCAN	Density-Based Spatial Clustering of applications with noise
DTW	Dynamic Time Warping
ED	Euclidean Distance
MA	Moving Averages
GF	Gaussian Filter

1 Introduction

The shortening of the product life cycles has been aggravated over the last few decades. SLC items are nowadays prevalent in several industries. Customers demand novelty in sectors such as fashion, technology, and other specific retail segments and, therefore, items remain on the shelves for shorter periods of time (Syntetos, A. A., Boylan, J. E., & Croston, J. D., 2005).

The reduced life cycle has implications on the forecasting process. Due to the fast-changing trends, intermittency, strong fluctuations, and lack of sufficient data, predicting future demand is a complex endeavor for SLC items (Afifi, A. A., 2020).

Forecasting demand for LLC and SLC products can be considerably different. While LLC products demand can be accurately predicted by traditional methods, SLC items present higher variability and more complex patterns, Relex Solutions, (2019), which require other forecasting techniques such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), Triana, M. J. B. (2012), and Support Vector Machine (SVM), Li, F. (2022).

The classification of product life cycle highly improves forecasting accuracy. Therefore, a fundamental advantage is found in the automatization and standardization of the SLC and LLC items identification procedure, mitigating the time and efforts allocated to analysis and leveraging the insights extracted from the sales data. Sales data analysis is a consuming task particularly in the retail sector where thousands of products are typically in the assortment.

1.1 Case study at Retail Consult

The present dissertation project was conducted at Retail Consult, more specifically in the planning unit of this company. Retail Consult is a Portuguese consultancy company, based in Porto. RC (Retail Consult) has a presence in seven countries and counts with over 250 workers, currently. RC operates in the B2B service market, providing strategical implementation, training and support on technological solutions oriented to retail.

Retail Consult operates in close partnership with Oracle, a large north American multinational focused on the production and commercialization of hardware and software. In this context, RC is specialized in Oracle products which are oriented to retail stakeholders. The Oracle Retail Demand Forecasting is one of the tools implemented in the **planning unit**. This tool offers a solution for demand forecasting based on statistical models that promises high forecast performance and reliability with reduced user interaction.

RDF requires users to define which products are SLC or LLC. This separation takes place due to the different analysis conducted for each of these item types. The identification and analysis of sales data is a procedure that requires deep understanding of the behaviors and indicators of the products. In the retail sector, this task is particularly difficult because of the soaring number of products in the assortment.

The current methodology to classify SLC and LLC items is based on business judgements, or the analysis of indicators widely associated with these types of items. The process is unstandardized, and limited research has been conducted to address this problem (Syntetos, A. A., Boylan, J. E., & Croston, J. D., 2005).

1.2 Objective and methodology

The main goal of the project is to improve the demand forecasting efficiency and performance, by creating a model to accurately classify SLC and LLC products. For that reason, historical sales were extracted from two major clients at Retail Consult, whose names will be omitted for confidentiality reasons. The first company represents a large Portuguese retail leader that detains a large supermarket chain across the country but also diversifies over multiple sectors such as shopping centers, telecommunications, and others. The second is a renowned Spanish company in the fashion industry with diverse locations across the globe.

With the resulting data set, clustering techniques are applied to group items according to the degree of similarity of sales behavior. This procedure is combined with data preparation and parameter optimization to model a classification framework that separates SLC and LLC items. The adequacy of the implemented approach is exhaustively analyzed and validated. An extensive demand forecasting analysis is conducted using Simple Exponential Smoothing (SES), Holt-Winters (HW), and Auto Regressive Integrated Moving Average (ARIMA) statistical models. The identified SLC class shows high forecasting errors and strong bias with these methods. To further evaluate the classified sample, indicators of multiple demand attributes are used and compared.

The project aims at achieving a SLC class characterized by strong fluctuations, intermittent demand, sell for few weeks, and provide poor forecasts, and an LLC class characterized by demand that is non-intermittent, regular over the long-term and accurately forecastable. This characterization aligns with the typical behavior of SLC and LLC items.

1.3 Dissertation structure

The present dissertation is divided in five additional chapters, organized as follows:

Chapter 2 is dedicated to an extensive overview of the relevant concepts such as product life cycle, item classification, clustering, demand forecasting, among others. As well as the methods approached for data preprocessing.

Chapter 3 describes in detail the current practice, its main drawbacks, the improvement opportunities, and obstacles found.

Chapter 4 details the proposed methodology and covers the model structure, data preparation, clustering, demand forecast, test design and model assessment.

In Chapter 5, the results of the framework implementation are presented and discussed based on the analysis of two experiments.

To conclude, Chapter 6 draws the main conclusions along with guidelines for future work.

2 Literature review

The present chapter aims to provide a general overview of the current understanding of product life cycle and in particular SLC products. Furthermore, the background and theory on the available forecasting techniques, analysis frameworks, evaluation metrics and clustering techniques adapted to time series are detailed. Section 2.1 introduces the product life cycle concept and the summary of research on SLC items' demand. Section 2.2 approaches the diagnosis indicators oriented to classify the demand forecast. The forecasting techniques and the evaluation metrics are introduced in Section 2.3. Section 2.4 approaches the clustering techniques and distance measures widely used for time series clustering. Section 2.5 describes preprocessing tools used for time series data. The chapter concludes with a summary in Section 2.6.

2.1 Product Life Cycle and SLC products

The life cycle of a product can provide key insights into its demand behavior and has become a concept of growing academic attention throughout the last few decades. The demand for an item goes through four main stages, historically recognizable, that make up the product life cycle. Introduction, Growth, Maturity and Decline. Levitt, (1965) provide a comprehensive analysis of the four PLC stages.

The product life cycle concept was proposed by Levitt, T. (1965) and refers to the period between the first launch and final withdrawal, during this period significant changes in the behavior of the product in the market take place, Komninos, I. (2002). In the introduction stage, the product is placed on the market and sales are usually slow while product awareness is not yet significant. If the item is successfully adopted by consumers, it starts the growth stage. At this point, sales take off and show increasing growth. Eventually, this growth starts to decelerate, and sales stabilize around a given level as the product enters the maturity stage. Some products may remain in this stage for most of their lifetime. Finally, and approaching the end of its life cycle, the product declines. Sales decrease as the product loses traction in the market and becomes obsolete **(**Levitt, T., 1965).

Short life cycle (SLC) products have become increasingly common in many sectors like fashion and technology. A typical demand curve for these products is characterized by rapid trend shifts and consists of transitory stages of growth, maturity, and decline coupled with seasonal variation caused by external market factors or occasional factors internal to the firm. The life cycles of SLC products are typically between three and six months Şen, A. (2008), but can range from a few months to a maximum of **3** years. SLC demand is characterized by scarcity of available data for future prediction (Kurawarwala, A. A., & Matsuo, H., 1998).

On the supply chain level, SLC items are particularly challenging. SLC requires rapid time to market and short pipelines to enable continuous replenishment during the life cycle. Spikiness and unpredictability in the demand of SLC increase the risk of obsolescence and lost sales, Aitken, J., Childerhouse, P., & Towill, D. (2003). The analysis and classification of product sales becomes important to determine the adequate forecasting approach for each item category.

2.2 Items classification

When conducting forecasting analysis there are several indicators that can be considered before applying forecasting methods. The analysis of historical sales employs data preparation techniques as well as demand and business attributes to extract insightful information.

This initial step is vital for selecting the forecasting technique and classifying the items. Besides its application for demand forecasting, these attributes are also important in other functional business tasks. The classification of sales volume, intermittency (p), and variability are relevant for inventory control. These classifications highlight interesting relationships in the data, detecting trends and demand behavior.

The ABC classification ranks items according to their importance for the supply chain, quantifying the impact of each item's sales for the overall company. Starts by sorting the items by sales (units or turnover) and calculate the cumulative sum. The groups are then created according to the *Pareto principle*, where 20% of the items that generate 80% of the returns are ranked as A products. Label A includes the items that profoundly impact the overall sales return, followed by B and C classes (Kheybari, S., Naji, S. A., Rezaie, F. M., & Salehpour, R., 2019).

The XYZ classification evaluates the level of variation experienced by the demand of each product. The level of variation is measured with the coefficient of variation of the sales. The coefficient of variation represents the ratio between the standard deviation and mean of the time series - *Equation* 1.

$$CV = \frac{\sigma}{\mu} \qquad Equation\ (1)$$

This method starts by computing, sorting, and applying the cumulative sum of the coefficients of variation. The X, Y and Z groups are then defined according to the distribution of the cumulative sum of CVs. X products typically have constant and smooth variation, while Y are associated to strong fluctuations with trend or seasonal patterns. Finally, Z items are considered irregular (Bulinski, J. *et al*, 2013).

The combined **ABC-XYZ classification** results in nine categories, classifying items in these two dimensions, sales contribution, and demand variation. AX items are associated to high sales value and smooth demand, whereas CZ products are marked by low sales contribution and sporadic demand (Pandya, B., & Thakkar, H., 2016).

The **Smooth**, **Intermittent, Erratic** and **Lumpy classification** further identify indicators to qualify the forecast ability of the demand. This framework evaluates two dimensions of the demand, the Intermittency (p) and Square Coefficient of Variation (CV^2). Items are considered intermittent if the p value is above the industry acceptance value of 1.32. On the other hand, regular demand items are considered those with CV^2 below 0.49. Items with values above this threshold are considered irregular (Mukhopadhyay, S., Solis, A. O., & Gutierrez, R. S., 2012).

Smooth items (S) express timely sales with stable variation. This category combines items that are non-intermittent and of regular demand. The Intermittent items (I) have stable variation of sales quantity, however, experience high intervals between sales. The intermittent category groups products that are intermittent and have regular demand. The erratic (E) category includes products with regular sales but with soaring demand variability, combining non-intermittent items of irregular demand. Finally, lumpy (L) demand is associated to intermittent products of irregular demand. The forecasting of these items is complex, and the accuracy is considered poor.

The forecast ability deteriorates over the classes S, I, E, and L. Where smooth items are considered the most forecastable, and lumpy associated with the poorest predictions.

2.3 Demand forecasting

Forecasting is a set of techniques that adds extensive value to a wide variety of applications. The current section refers to forecasting methods for demand time series. As well as performance metrics used to evaluate the forecast errors. A time series is a chronologically ordered collection of sales data.

Demand forecasting techniques can be of two subgroups, qualitative or quantitative methods. Qualitative forecasting requires judgements, whereas quantitative methods rely on statistics and data-driven procedures. The project focuses on quantitative forecasting methods to evaluate the forecasting environment of the products classified as SLC and LLC.

Under the quantitative techniques, statistical forecasting methods present a wide range of algorithms with varying complexity that assume the dependency of future values with past observations. Multiple models have been proposed for statistical prediction, **Simple Exponential Smoothing** (SES) one of the simpler models proposed is not sensible to quickly changing trends. The triple exponential smoothing, widely recognized as **Holt-Winters** (HW), provides additive and multiplicative variations, and considers trends and seasonal patterns. The **Auto Regressive Integrated Moving Averages** (ARIMA) method approaches forecasting by describing correlations in the data, comparing the original time series with lagged versions. Other approaches such as SARIMAX, regression methods and Croston's are also considered statistical methods (Nenni, M. E. *et al.*, 2013).

Besides statistical techniques, other *machine learning* approaches have been proposed for forecasting of time series. Some of the methods applied for this purpose are Neural Networks (NN), Random Forests (RF) and Gradient Boosting Machine (GBM) (Garg, R., & Barpanda, S., 2022).

(Triana, M. J. B., 2012) applies two forecasting methods to predict SLC items demand: multiple linear regression (MLR) and support vector regression (SVR). It applies time series clustering prior to demand forecasting for accuracy improvement. (Li, F., 2022) proposes a new method for forecasting and supply chain optimization of a medical consumable SLC item, during the COVID-19 pandemic. The paper uses a support vector machine (SVM) algorithm to predict the demand of short life cycle products based on factors such as price, season, promotion, industry status, historical sales data, and channel type. The paper achieves improvements on the forecasting accuracy, efficiency of the supply chain design and environmental impact of logistics operations.

For the presented methodology, three classical statistical methods were selected for forecasting: SES, HW, and ARIMA.

2.3.1 Simple Exponential Smoothing

Simple Exponential Smoothing (SES) uses the exponential window function to forecast time series data. The method assumes that the future behavior will be like the recent past.

The model assigns exponentially decreasing weights to distant instances and produces forecasts that are a linear weighted sum of past observations. SES model parameter α represents the smoothing parameter.

$$f_t = \alpha * Y_{t-1} + (1 - \alpha) * f_{t-1} \qquad Equation\ (2)$$

In *Equation* 2, α is the smoothing parameter, Y_{t-1} is the previous value of the time series and f_{t-1} the previous prediction.

SES is a simple forecasting technique that performs well on data without significant trend or seasonality for short term forecasting. Since other parameters are not considered, SES may not be suitable for complex trends or seasonal patterns. However, it can predict accurately simple time series with stable demand (Ostertagova, E., Ostertag, O., 2012).

2.3.2 Holt-Winters

Holt-Winters (HW), also known as triple exponential smoothing, is an extension of the exponential smoothing method that can handle stable and predictable trends and seasonal patterns. This method requires 3 parameters for level, α, trend, β, and seasonality, γ. Furthermore, the model features variations of two types: additive or multiplicative, for both the trend and seasonal components.

HW is suitable for time series with consistent patterns of level, trend, and seasonality. However, this method may perform poorly for irregular data with fast changing patterns. In that case, the approach may not be suitable for the series at hand.

$$L_t = \alpha * (Y_t - S_{t-m}) + (1-\alpha) * (L_{t-1} + T_{t-1}) \qquad Equation\ (3)$$
$$T_t = \beta * (L_t - L_{t-1}) + (1-\beta) * (T_{t-1}) \qquad Equation\ (4)$$
$$S_t = \gamma * (Y_t - L_t) + (1-\gamma) * S_{t-m} \qquad Equation\ (5)$$

$$F_{t+h} = L_t + h * T_t + S_{t-m+h} \qquad Equation\ (6)$$

Equations [3-6] formulate the additive HW algorithm. In this model, three calculations are made to estimate the level, trend and seasonal components of the time series using parameters α, β, and γ. The forecast is then obtained by summing the level value to the trend value multiplied by the forecast horizon, h, and seasonal lag component. The multiplicative variations of HW follow parallel logic to the additive model and can handle multiplicative behavior for both trend and seasonal components (Kalekar, P. S., 2004).

2.3.3 ARIMA

When compared to Exponential Smoothing methods, ARIMA models provide a different approach to time series forecasting. This technique aims to find autocorrelations, i.e., correlation with the lagged series, and assumes non-stationarity.

ARIMA requires the definition of three parameters: Auto Regressive, p, Integrated, d, and Moving Average (MA), q. The auto regressive component models the dependence of the current value on past instances where p indicates the number of lagged terms to be included. The integrated component is controlled by parameter d and represents the degree of differencing required to make the series stationary. Finally, the MA part incorporates the dependency between the observation and the residual errors of the moving average with window q, calculated with the lagged observations. The ARIMA model is defined in *Equation* 7.

$$\left(1 - \sum_{i=1}^{p} \phi_{L^i}\right)(1-L)^d X_t = \delta + \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t$$

Equation (7)

Where L is the lag operator, ϕ are the coefficients of the AR term, θ the coefficients of the MA term, p, d, and q the parameters of the ARIMA model and ε the residual values.

This method can predict several types of time series, whether with trend, seasonality, or irregular patterns. However, may require resources and expertise for parameter selection (S.L. Ho, M. Xie, 1998).

2.3.4 Error metrics

To evaluate forecasting performance, metrics are usually calculated and evaluated. Several accuracy measures have been widely used to quantify errors. Some of the most popular metrics are the Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Error (ME), and the Symmetric Mean Absolute Percentage Error (SMAPE).

These indicators evaluate various aspects of the forecast's residual error. MSE measures the magnitude of the errors, penalizing larger ones by squaring the difference between the forecast and actual values for each observation *Equation* 8. This measure is extremely sensitive to outliers. On the other hand, MAE represents the absolute difference of the forecast and actual values, treating all errors equally *Equation* 9. In contrast with the previous metric, MAE is more robust to extreme values. ME represents a measure of bias *Equation* 10., i.e., systematic error that reflect a tendency to over or under-forecast. Lastly, SMAPE is a symmetric metric (equally penalizes under and over forecasts) based on percentage absolute errors *Equation* 11. This performance indicator is preferable to the more traditional Mean Absolute Percentage Error (MAPE), since MAPE produces unreliable values when actual instances are close to 0, overestimating the magnitude of these errors (Petropoulos, F. et al., 2022).

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} (f_i - a_i)^2 \qquad Equation\ (8)$$

$$ME = \frac{1}{n} * \sum_{i=1}^{n} f_i - a_i \qquad Equation\ (9)$$

$$MAE = \frac{1}{n} * \sum_{i=1}^{n} |f_i - a_i| \qquad Equation\ (10)$$

$$MAPE = \frac{100}{n} * \sum_{i=1}^{n} \frac{|f_i - a_i|}{(|f_i| + |a_i|)/2} \qquad Equation\ (11)$$

Due to the normalization of the time series, the forecast MSE calculated for the scaled sales may have a lower relative amplitude when compared to the one calculated in the original scale. This is caused by the square of the errors, which penalize larger values. Despite not representing the actual magnitude of the MSE for the forecast in the original scale, this metric allows the comparison of the forecast performance among all scaled time series in the data set.

Since MAE, ME, and SMAPE operate absolute values, these metrics are not as affected by the scale of the data and incorporate the true relative amplitude of the errors.

The forecasting error can be reduced by applying clustering techniques to group items according to sales behavior and adapting the forecasting process to the characteristics of each cluster. For the scope of the project, SMAPE was selected since it provides a percentage symmetric error measure. ME was also selected to measure forecast bias.

2.4 Clustering of time series

This section comprehends an overview of the clustering techniques applied to time series data. It starts by covering the distance metrics used to compare time series data and serve as ground for clustering. Then the clustering algorithms are comprehensively introduced, and the evaluation metrics described.

Clustering aims to find underlying structures in unlabeled data sets, assigning data points to clusters by minimizing intra-cluster similarity and maximizing inter-clusters dissimilarity. Clustering is required when labeled data is not available, independently of the type of data in analysis. Most of the past work developed in this field was oriented to static data, i.e., data whose features do not change significantly over time, Aghabozorgi, S., *et al*, (2015). However, the growing ability to store time series data for extended periods urged the need to efficiently analyze dynamic data, resulting in increased interest on time series clustering techniques for multiple applications in supply chain, finance, bioinformatics, among others, in the last decades, T. W. Liao (2005) and López-Oriona, Á., *et al*, (2022).

Time series clustering is a fundamental problem of machine learning and is mostly employed to discover meaningful patterns in time-series data. A time series is a sequence of continuous real-valued elements, registered chronologically on a fixed period (daily, weekly, quarterly, yearly, etc.). However, a time series can also be modeled as a single object, Aghabozorgi, S. *et al*, (2015). Clustering of sales time series data in the retail sector has been proposed for the identification and estimation of seasonal patterns in Kumar, M., *et al*, (2002).

(He, Y., Li, J., Zhu, M., & Xu, D., 2018) explore clustering algorithms to identify the life cycle of microblog hot topics that reflect public opinion. This paper proposes the application of clustering algorithms to transform and classify time series data of hot topics. It uses the SOM (self-organizing map) algorithm to cluster 75 life cycle curves into five life cycle stages.

(Novia, C., Santoso, I., Soemarno, S., & Astuti, R., 2020) classify the life cycle stage of a food product (apple chips) sales across 31 SMEs (small and medium-sized enterprises) in Indonesia. The paper proposes the application of k-means clustering for sales data to classify apple chips SMEs in three categories: introduction, growth, and maturity.

Clustering algorithms can be segmented into three main approaches: partitional, density based, and hierarchical. Partitional methods like the k-means divide the data space into *k* subsets, assigning each observation to one and only one cluster. Density-based methods, however, define clusters as dense regions separated by sparse regions in the set. Finally, hierarchical clustering generates nested hierarchies represented by a dendrogram. The formation of clusters at the distinct levels is then performed by merging (agglomerative) or dividing (divisive) hierarchy levels until a stopping criterion is met (Jain, A. K., *et al.*, 1999).

2.4.1 K-means

The k-means algorithm is an effective partition-based clustering technique that can be applied to identify patterns in time series.

It starts by randomly selecting k data points from the set as the initial centroids. Then, a precomputed distance metric is used to assign the remaining instances to the clusters. At each iteration, the k centroids are recalculated as the mean distance to each centroid, Jain, A. K. (2010).

$$f = \sum_{k=1}^{K} \sum_{x \in C_k} d(x, m_k)^2 \qquad Equation\ (12)$$

In this technique, the objective function of $Equation$ 12 is minimized to assign data points to the K clusters. Where $d(x, m_k)$ is the distance between the series x and the centroid m_k (Hartigan, *et al.*, 1979).

This approach is simple yet effective for grouping data into homogeneous clusters when the number of clusters, k, can be determined a priori.

2.4.2 DBSCAN

DBSCAN is a density-based algorithm that discovers clusters of arbitrary shape. The method can handle data with varying density and adapting to its underlying structure and was proposed by Ester *et al*, (1996).

This model defines clusters as dense regions separated by sparser regions. **The key idea is that each neighborhood radius defined by the *eps* parameter around a given observation must contain a minimum number of objects (*MinPoints*) to assign it as a core point**. Density-based models offer notable advantages over partition-based algorithms since they model the density structure of the data.

The DBSCAN algorithm proceeds iteratively through the unseen observations and assigns them as core or border points. A core point is one that has a minimum number of neighbors within an *eps* radius. Instances that do not meet these requirements are assigned as borders or noise points. The clusters are formed by expanding the neighborhood of both core and border points.

$$N_{eps} = \{\, q \in D \;/\; dist(p, q) < eps \,\} \qquad Equation\ (13)$$

N_{eps}, is the neighborhood of an arbitrary point p defined in *Equation* 13, where D represents the data set objects, and *eps* the neighborhood radius. If the neighborhood of a point P contains more than the minimum number of points (*MinPoints*), the observation is called a core point, *Equation* 14.

$$N_{eps}(P) \;>\; MinPoints \qquad Equation\ (14)$$

The algorithm checks the neighborhood of each object in the set and if the neighborhood contains more than *MinPoints*, a new cluster with p as a core object is created. The process ends when no new object can be added to any cluster (K. Khan *et al.*, 2014).

2.4.3 Dynamic Time Warping

When it comes to distance metrics used for comparing time series, The Euclidean and DTW distances are widely applied. Dynamic time warping (DTW) is a distance metric that allows the comparison between time series, whose patterns may be similar but unsynchronized in time.

DTW is used to calculate minimum distances between time series. To achieve it, the algorithm searches for the optimal alignment between each pair of time series by locally distorting the temporal axis of one in relation to the other. This technique calculates the Euclidean distance *Equation* 15 between each pair of time series across all alignments to find the minimum distance between them, which occurs in the optimal alignment. This allows for non-linear relationships between time series to be considered in the calculation of the distances, *Fig*. 12.

$$ED_q(x, x') = \left(\sum_{(i,j) \epsilon\, \pi} d(x_i,\ x'_j)^2 \right)^{1/2} \qquad Equation\ (15)$$

To compute the DTW distance *Equation* 16, each alignment, π, represents a sequence of *n* pairs of (x_i, x_j*')* series, considering the set of possible solutions, *A(x,x')*. The euclidean distance is calculated between each alignment and minimized to find the optimal distance.

$$DTW_q(x, x') = \min_{\pi\, \epsilon\, A(x,x')} \left(\sum_{(i,j) \epsilon\, \pi} d(x_i,\ x'_j)^q \right)^{\frac{1}{q}} \qquad Equation\ (16)$$

In spite of the multiple advantages of calculating the distances between series using DTW, this method presents sizable computational cost when compared to the euclidean distance. As a result, large data sets require dimensionality reduction for better efficiency (Tiwari, *et.al.*, 2017).

2.4.4 Cluster evaluation metrics

The cluster performance metrics are based on similarity measures within and between groups. These indicators evaluate the degree of resemblance between points assigned to the same cluster and the separation between points assigned to the competing clusters. The referenced metrics used for clustering evaluation are the **Silhouette Score** and Davis-Bouldin index. The latter is measured as a ratio between the intra-cluster (S_i) and the inter-cluster ($M_{i,j}$) distances, *Equation* 17 (Petrovic, S., 2006).

$$DB = \frac{1}{n} * \sum_{i=1}^{n} \max_{i \neq j} \frac{S_i + S_j}{M_{i,j}} \qquad Equation\ (17)$$

The Silhouette Score provides a measure to quantify the clustering quality of a given model. This score is calculated for each data point and averaged to achieve the overall Silhouette Score, *Equations* [20, 21]. This metric evaluates the cluster attribution on two levels, the within-cluster distance, and the between-cluster distance. Where a_i measures the compactness of each cluster, i.e., how similar the data points within each cluster C_i are - *Equation* 18. And b_i measures the degree of dissimilarity between the C_k clusters *Equation* 19.

$$a_i = \frac{1}{|C_i| - 1} * \sum_{j \epsilon C_i,\ i \neq j} d(i, j) \qquad Equation\ (18)$$

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} * \sum_{j \epsilon C_k,} d(i,j) \qquad Equation\ (19)$$

$$SS_i = \frac{b_i - a_i}{\max(a_i, b_i)} \qquad Equation\ (20)$$

$$SS_T = \frac{1}{n} * \sum_{i=1}^{n} SS_i \qquad Equation\ (21)$$

The interpretation of the silhouette score follows specific guidelines. Where higher results are associated with better clustering performance. The silhouette score ranges from -1 to 1, where values close to 1 indicate a sample that shares high similarity within and dissimilarity between clusters. A SS close to 0 may suggest that the sample has disperse decision boundaries between neighboring clusters. Furthermore, values near -1 indicate misclassification of the sample (K. R. Shahapure, C. Nicholas, 2020).

2.5 Time series preprocessing

Applying forecasting techniques directly to raw data can result in poor performance. Therefore, there is an extensive set of methods for preprocessing time series data. The main tasks of this procedure are outlier treatment, noise smoothing and normalization.

Across these processes, some transversal tools are considered to analyze each component of the data anomalies. Firstly, smoothing techniques are employed to isolate random variation and cancel it in the data. This generates smoothed trends that represent the underlying structure of the series. Then, normalization or standardization is usually required for machine learning algorithms.

2.5.1 Smoothing techniques

Moving Average (MA): Moving averages are key tools when working with time series data. This smoothing method is a simple weighted sum dynamically calculated over historical sales data. For each point, the mean of the previous *n* sales is calculated and used to replace the actual value. This results in a delayed smoothed representation of the original time series.

$$MA_t^n\ (Y_t) = \frac{1}{n} * \sum_{i=0}^{n} Y_{t-i} \qquad Equation\ (22)$$

Where Y_t is the time series and $MA_t^n\ (Y_t)$ the smoothed value of the observation t obtained with the MA of window *n*, in *Equation* 22.

This method is a simple manner of regularizing time series data that expresses prevalent random variation. The window parameter of the moving average defines the number of previous weeks to consider in the calculation of the $t + 1$ weighted value. It is important to notice that the method of *Equation* 1 requires *n* previous observations to estimate each observation, therefore, the first *n* observations of the series are not estimated but are used to compute the subsequent *n* MAs. Furthermore, the parameter *n* has a significant impact on the lag of the smoothed series, higher *n* values lead to longer lags and lost data, **in** Raudys, A. *et al*, (2013).

Gaussian Filter (GF): The gaussian filter is a smoothing technique widely used for image processing. However, its properties permit leveraging the noise smoothing procedure for sales time series, Roberts, S., et al. (2013). Gaussian filter smooths data by weighing nearby observations based on a Gaussian distribution. The smoothed array is obtained by convolving the time series with a Gaussian kernel, a function based on the Gaussian distribution, *Equation* 23.

$$GF(Y_t) = \frac{1}{2\pi\sigma^2} e^{\frac{-(t^2 + Y_t{}^2)}{2\sigma^2}} \qquad Equation\ (23)$$

With Y representing the time series, the parameter σ is the window (standard deviation) of the gaussian kernel and controls the degree of smoothing. Where higher σ parameters produce wider smoothing of the series, in C. Lopez-Molina *et al*, (2013).

This technique results in centered smoothed time series that conserve sharp edges in the data while removing random variation. The method is dependent on the knowledge of future values at each instance. Therefore, different samples taken from the same time series will result in different smoothed values for the equivalent observations.

2.5.2 Scaling

Normalization techniques are applied to scale down data for later usage. This stage plays a significant role in machine learning algorithms. There are many normalization techniques such as **Min-Max**, Z-score, and Decimal scaling (Patro, S. G. O. P. A. L., & Sahu, K. K., 2015).

The Min-Max scaling technique of *Equation* 24 applies linear transformation on the original range while keeping the relationship among original data. This method can fit the data in specific pre-defined boundaries.

$$x_{scaled} = \frac{x - min(X)}{max(X) - min(X)} \qquad Equation\ (24)$$

This normalization method produces time series scaled in [0, 1], with 0 representing the lowest and 1 the highest sales registered for time series X.

2.6 Summary

To conclude this chapter, a summary of the literature review is presented. Referring to the main takeaways of the research conducted.

Smoothing methods such as Moving Averages (MA), SES and HW are applicable to steady trends for short-term forecasting. For these techniques, abrupt changes of trend lead to systematic error in the forecast, hereby measured by the Mean Error (ME). With that said, the application of these traditional methods for SLC products has severe drawbacks. Therefore, the forecast performance of these techniques indicates whether an item is SLC or LLC.

The performance metrics are designed to evaluate errors under diverse perspectives. SMAPE is a percentage absolute measure to quantify the relative variation of the errors. This indicator provides insights on the relative magnitude of the errors. ME is the simple average of the errors. This indicator measures the bias of the generated forecasts, where values normally distributed with mean 0 indicate un-biased predictions.

The application of clustering algorithms for time series has become of high relevance over the last decades. Adaptations of these methods were proposed for identifying patterns in sales data.

K-means and *DBSCAN* are techniques applicable for time series clustering. *K-means* finds groups as homogeneous partitions of the data, while *DBSCAN* assigns as clusters the dense regions separated by sparser ones.

Dynamic Time Warping (DTW) is a distance metric for time series that can be employed for the *k-means* and *DBSCAN* algorithms. *DTW* finds the optimal alignment between data points to find the minimum distance between the distorted series.

Clustering techniques cannot be evaluated by comparing the labels produced with actual values since these are unknown. To measure cluster performance, it is necessary to measure similarity among data and calculate metrics like the *silhouette score* or *Davis-Bouldin index*.

Preprocessing is a step of foremost importance when dealing with sales data. *Moving averages* are widely used in forecasting, this technique produces a delayed smoothed representation of the time series. *Gaussian Filter* is another smoothing technique able to obtain centered smoothed versions of the time series by Gaussian approximation.

The normalization or standardization procedure is required for clustering algorithms and scaled comparison of results.

3 Problem description

Short life cycle products are more than ever preponderant in the retail sector. Consumer patterns have become increasingly transient in the last few decades. As customers pull constant innovation in a wide variety of lines, retailers must meticulously manage their portfolios, frequently introducing new products and streamlining failures to accompany market shifts. The replenishment of SLC items requires accurate predictions for efficient planning, especially when lead times are sizable. However, this is usually hard to accomplish with traditional statistical forecasting methods when it comes to SLC products. Several alternative methods have been proposed to improve the accuracy for these items' demand forecast.

The forecasting process is an overly complex task when it comes to SLC products. The lack of representative historical sales creates an obstacle for most traditional demand forecasting methods that require sized data for fitting. Moreover, the abrupt trend changes and high volatility associated with SLC products add another layer to this problem. Even though forecasting is a crucial component to optimize and leverage parallel business tasks, demand prediction is the most common application among most solutions requested in consulting projects at Retail Consult.

Forecasting solutions, like Oracle Retail Demand Forecast (RDF), offer statistical models that apply multiple techniques according to product life cycle and other user-defined hyper parameters and overrides. The historical data is preprocessed, trained, and then tested with adequate methods. The forecasts are finally produced and used across the business. Besides its many features, RDF does not provide a solution for **automatic identification of the life cycle** (SLC, LLC). This information is currently required as input and users must manually classify the products, which is a demanding task in retail, marked by soaring numbers of SKUs.

Limited research has been done on the identification of short life cycle items, even though their adequate classification is key for accurate and timely demand forecasting. In this context, the project proposes a solution for classifying products according to their life cycle (SLC, LLC). Resulting in a framework *python* application, further discussed in Chapter 4. The current chapter intends to state the identified problem and its requirements. The first section details the current processes followed to classify and forecast the demand of SLC and LLC by retailers. Then the proposed approach is briefly presented as well as the faced issues and respective implemented solutions.

3.1 Current approach to SLC products

The current methodology implemented at Retail Consult for demand forecasting is based on the RDF solution. This application offers multiple features. Selects the best model based on the demand characteristics and automatically evaluates several methods among a wide selection of techniques like **Simple Exponential Smoothing** (SES), **Holt-Winters** Exponential Smoothing (HW), Additive and Multiplicative, **Autoregressive Integrated Moving Averages** (ARIMA), Holt Linear (H), Croston's Intermittent, etc. Additionally, RDF allows combining the output of multiple methods for more robust forecasts that offset overfitting.

The forecasting process consists of i.) data preparation, ii.) parameter estimation, and iii.) forecast. During data preparation, stock-outs, promotions, and outliers are identified and treated to improve the reliability of the data. For LLC, data is usually reliable, so anomalies are corrected. On the other hand, SLC products data tend to be unreliable over short periods, so these anomalies are omitted. Then, parameter estimation aims to find the optimal set of parameters for forecasting methods at all escalation levels, i.e., groupings of items and locations usually tied to explicit hierarchy levels. At last, the estimated parameters are applied, and each model is individually optimized and evaluated. The forecasts are produced with best-fit models or blended models.

Despite the multiple built-in functionalities of the base RDF, Retail Consult implements customized solutions tailored to specific requirements and contexts. An issue commonly arises when it comes to separating the SLC from the LLC items. Retailers face up to multiple thousands of SKUs to forecast demand, and dedicating attention to the analysis of all these product's sales individually is not feasible.

Currently, SLC items are classified considering qualitative and quantitative factors analyzed empirically or data driven. By visualization, SLC products demand is characterized by i.) **rapid growth** followed by **abrupt decline** with few to no maturity, presenting transient, non-stationary and non-linear patterns. SLC products can sell from a couple of weeks up to three years, but overall, ii.) remain by a **limited time in the market** before becoming obsolete. iii.) **High demand volatility** is also associated with SLC items (CV^2 above 0.49) as well as iv.) **high intermittency** (p above 1.32), in Section 2.2. The accurate forecasting of these items requires deep knowledge on the characteristics and behaviors of the demand, and usually calls for **sophisticated predictive models** mentioned in Chapter 2. On the other hand, LLC items show linear behavior over prolonged periods of time, permitting accurate sales forecasts with traditional methods (Triana, M. J. B., 2012).

The current process of life cycle classification is unstandardized and varies from company to company or even across different data analysts. The current paradigm lacks an automated and standardized procedure to classify products as SLC or LLC.

3.2 Proposed solution

The solution implemented tackles the necessity of an automated and unbiased framework to identify and handle SLC. The model was designed for users with various levels of expertise that aim to separate SLC and LLC products within the assortment for forecasting or exploratory purposes.

The proposed solution assumes that SLC and LLC products possess inherent patterns which are interpretable, similar within each category and different between categories. The project approached the identification of SLC products using clustering techniques adapted to time series data. Clustering is a widely used approach to find natural relationships in the data and has applications in diverse research areas and contexts that require working with unlabeled data. This approach can group products according to intrinsic characteristics and, if combined with classification, detect SLC and LLC.

Sales data in the retail reality is usually flawed. The analysis of this data requires extensive efforts to extract valuable insights. With the automated approach, multiple issues arise when comparing demand patterns and interpreting forecast results to classify products as SLC or LLC. The purpose of the following paragraphs is to expose and summarize the obstacles faced as well as the actions considered to mitigate their effects during the project.

Time series comparison: Comparing time series is a key part of the proposed framework. The demand of a product can be characterized under several perspectives, considering the demand components (level, trend, and seasonality), statistics (standard deviation, mean, max, etc.) or business metrics (sales volume, coefficient of variation, intermittence, sales range, etc.). To understand the behavior of SLC products that are clearly distinguishable from LLC products, the proposed analysis focuses on learning the demand components and patterns for these items, disregarding the absolute sales volume level. Furthermore, the error obtained with the implemented forecasting methods requires a common ground when comparing products forecast performance. With that in mind, the time series were individually normalized, on a scale ranging in [0, 1], so that 0 represents the actual absence of sales and 1 the sales peak of each time series. Combined with other data preparation processes: aggregation, extreme values treatment, and noise smoothing.

Evaluate model performance: Evaluating cluster performance implies measuring similarity within and between groups. This is usually performed by analysis of specific metrics like silhouette score or Davis-Bouldin index. These metrics evaluate the clustering performance but are not direct measures of its adequacy to the problem at hands. The framework presented relies on unlabeled data to classify products as SLC or LLC, not knowing exactly which product belongs to each category. This constitutes a challenge when it comes to evaluating the performance of the framework's classification approach as accuracy metrics require target data to be calculated. To contour this obstacle, the evaluation approach considered performance metrics to compare and interpret the forecasting environment for each group. Parallelly, the characteristics of SLC products referenced in the literature were cross-checked with the classification.

Distance Metric: Clustering implies calculating a distance measure between data points. To compare time series data, distances are usually computed with the mean Euclidean distance between each pair of time series. This metric can capture similar patterns that occur in the same temporal alignment, yet it is unable to identify them when they occur in distant temporal periods. Dynamic Time Warping (DTW) distorts the relative temporal axis to calculate the distance between time series in the optimal alignment. This means that the similarity is considered regardless of the original temporal alignment, comparing patterns that occur in different points in time. Despite the broad advantages, DTW is computationally costly, and its efficiency is limited to small and medium sized data.

Escalation: Data granularity plays a significant role in sales data analysis and prediction. Aggregation is the process of grouping sales according to product and location hierarchies. This technique controls the granularity of the sales data and in some situations leads to more robust representations of the demand. The analysis of the product life cycle was generalized to all locations' sales at SKU level. This representation of the data provides a holistic view of the sales registered by each SKU without considering location factors.

Sampling: The analyzed set contains historical sales data of 20060 SKUs, making computing the DTW distances for the complete set impracticable. Simple Random Sampling (SRS) is a sampling technique that randomly selects observations from the complete set. It conserves the data distribution of the original data, depending on the sample size and normality of the values, but falls short when representing minority classes in imbalanced data. Stratified Sampling (SS) on the other hand samples observations from each group in the data, which leads to representative presence of instances from all strata.

Extreme Values: Extreme values are accountable to stock-outs, promotions, and outliers. These events result in sales that deviate significantly from the trend. Extreme values aggravate the data's reliability and compromise conclusions. It is important that these observations are identified and treated before the analysis.

Noise: The presence of random variation confuses meaningful patterns in the data. Smoothing is a widely used approach to isolate the sales from noise. Moving Average (MA) is a simple yet effective way to meet this end. It smooths data by calculating the average of the *n* previous observations. Gaussian Filter (GF) on the other hand smooths data by approximating it to a gaussian kernel of standard deviation σ. MA is simpler and has lower computational cost when compared to GF but eliminates n observations when calculated and is less sensible to sharp edges in the series. GF was used for clustering and MA for forecasting purposes.

Model optimization: Most statistical and machine learning algorithms depend on the definition of multiple parameters. The choice of these values impacts the generated results, so model optimization is a fundamental endeavor. In general, optimization is achieved by training models with different parameter combinations and selecting the best performing model.

The proposed approach clusters time series relative to sales of 1000 SKUs sampled from a set that integrates historical data registered over the last three years. The clusters are then classified according to the number of spikes (local maxima) in the product time series. Lastly the approach is thoroughly evaluated. Three forecasting methods (SES, HW, ARIMA) are implemented to assess the forecasting environment for each class. Furthermore, other indicators were computed to further analyze the class attributes. The final framework was validated by comparing the prediction of the generated classes with the linear combination of SLC indicators, assuming industry thresholds.

4 Methodology

The current chapter aims at detailing the proposed approach to classify SLC and LLC products using historical sales to fit clustering models. In Section 4.1, the proposed solution is thoroughly detailed. Further in Section 4.2, the framework and data structures are presented. The implemented methodology begins in Section 4.3 with data preprocessing. Section 4.4 describes the process of clustering and classifying products based on sales data. In Section 4.5 the demand forecasting of the classified products is conducted and evaluated. Section 4.6 describes the evaluation procedure for the proposed classification model. Across the sections the implementation is detailed, and the required algorithms and indicators are introduced.

4.1 Proposed approach

The proposed framework previously summarized in the last chapter aims to classify products as SLC or LLC based on historical sales data. The model classification capability is evaluated by forecasting error metrics resulted from traditional forecasting methods as well as the estimated demand indicators for the SLC and LLC classified items.

The classification of time series is an extensive procedure, sales data require precise data preprocessing to be reliable, and optimizing the distances between time series is a costly process. The clustering of large data containing product sales time series is a complex task. Adjusting the scales of the time series is essential for these machine learning algorithms as they are sensitive to large values, so data preprocessing becomes necessary for model interpretability and robustness. The choice of clustering algorithm considers performance metrics and inspection of the spatial dispersion between data points. Two options were considered: **k-means** and Density-Based Spatial Clustering of Applications with Noise (**DBSCAN**).

To prepare the sales data, the proposed solution starts by aggregating sales to the SKU and company level. This process follows the product and location hierarchies when applicable. Then the extreme values are identified as observations that show extreme deviation from the trend component. After outlier treatment, the noise is reduced by application of Gaussian Filters or Moving averages.

After the described preprocessing, the data is clustered using Dynamic Time Warping (DTW) distances between time series. This process results in groups of products characterized by their demand behavior. The clustered products are then classified as SLC or LLC, considering the number of sales spikes within the time series and the selling range. SLC and LLC clusters were considered the ones with less and more average spikes, respectively. Since LLC products are typically more stationary and sell during longer periods, the number of local maxima observations in the smoothed time series are expected to be higher when comparing with SLC.

The diagram in *Fig.* 1 describes the framework of the proposed approach. Presenting the main components and procedures implemented as well as the algorithm workflow. Starting with raw sales data and producing classified labels for the products in analysis. The framework for the implemented solution to classify SLC and LLC products is based on three main components: preprocessing, clustering, and classification.

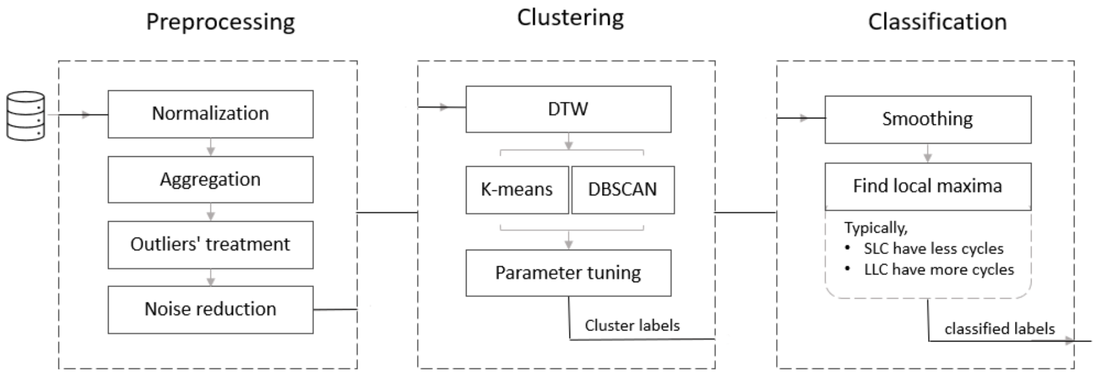

Fig. 1 – Framework of the proposed solution to identify SLC and LLC products.

Then the classification procedure is evaluated. Simple Exponential Smoothing, Holt-Winters and ARIMA statistical methods are implemented and optimized to predict future sales of the products. A robust representation of the model's performance is key to correctly assessing its forecasting accuracy in the analysis context. For that matter, the tests were applied with cross-validation across different sections over a user defined test window. The number of future values to be forecasted are also inputted by the user. Overall, the SLC items are poorly forecasted by these methods whereas LLC are accurately forecasted

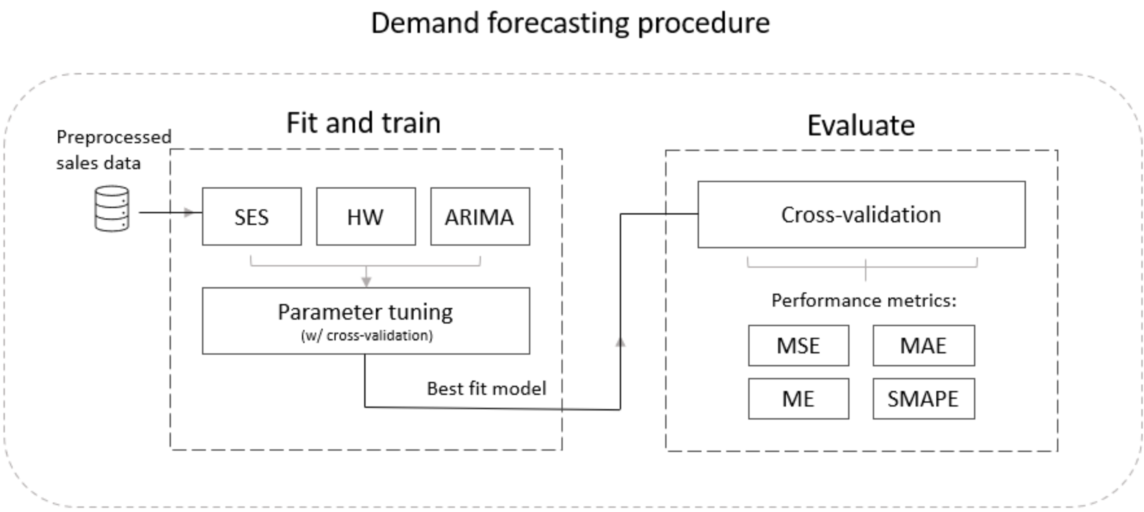

Fig. 2 – Framework to predict future demand and evaluate the forecasting performance.

Fig. 2 outlines the forecasting procedure endorsed to evaluate the classified labels. The preprocessed sales are trained with SES, HW, and ARIMA forecasting methods and evaluated with cross-validation calculating MSE, MAE, ME, and SMAPE accuracy metrics.

The presented analysis was performed under both short and medium terms, with test window of 16 weeks, and forecast windows of 1 and 8 weeks, respectively. Other demand attributes such as sales volume, coefficient of variation and intermittency were computed to support the analysis.

4.2 Framework Structure

The implementation of an application to support decisions that rely on the classification of products as SLC and LLC requires designing the workflow of information and user interactions. Such a solution must balance the customizable accessibility to the different components of the application without sacrificing its usability and interpretation.

Users interact with the application through the interface and dashboard, the internal processes are encapsulated for user-friendly simplicity and to mitigate corruption risk. In the interface, users import the required data into the application. The model requires sales data coded at the lowest granularity level and with a date column in acceptable formats. Product and location hierarchies are necessary when escalation of the data is required for the analysis.

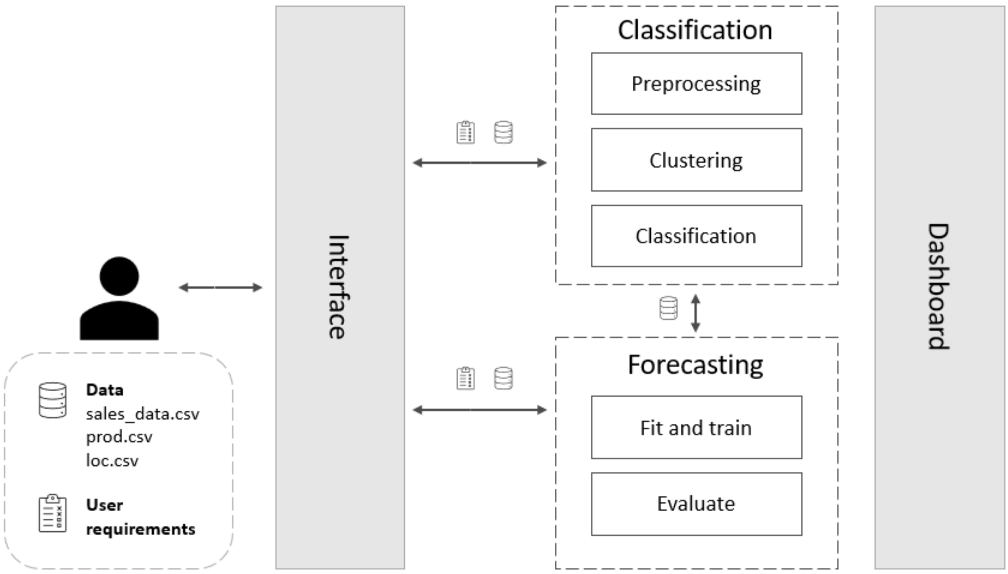

Fig. 3 – Representation of the user interactions and data workflow.

Despite being encapsulated, the framework components can be accessed and customized by users. The model assumes default settings for most options. However, users can define requirements for both classification and forecasting components as well as dashboard settings. For classification purposes, the model accepts hyper parameters to customize the preprocessing of data, clustering procedure and classification. Whereas for forecasting, model and evaluation options are also customizable. The classification and forecasting components bridge the communication of the sales data, where forecasting is performed only for the labeled products resulting from classification. Or for user defined samples.

The dashboard displays the performance of the model, evaluating it under three perspectives: clustering, classification, and forecasting. The clustering performance is analyzed by the silhouette score indicator, the spatial representation of the cluster points is also displayed visually in the dashboard. Under the classification evaluation, the business indicators are calculated for the clusters data points, validating the estimated characteristics with the current research on SLC products demand and industry assumptions. These indicators and a sample plot from each classified group are shown. Finally, the forecasting results are also presented with indicators and boxplots of the measured metrics’ distributions.

The implementation of the application was conducted in *python*. *Python* environments count with extensive libraries to handle multiple machine learning and time series concepts. The application is based on a variety of libraries for specific purposes and model tasks, explored over the following sections. The user-customizable variables for tailored analysis under the proposed framework are listed and described below, while these factors are optional and assume default settings when no user-defined option is provided. The input variables control distinct components: sampling, preprocessing, clustering - classification, and forecasting.

Sample size & **Sampling method**: Sampling takes place when the original sales data includes a soaring number of products. The framework allows the sampling of the complete data set with a customizable technique and sample size. Users can select stratified sampling (SS) based on a default threshold of the selling period, or simple random sampling (SRS) from the pool of candidate products. The size of the sample is customizable; however, the proposed analysis considered samples of 1000 products with sales aggregated at the SKU and company levels, i.e., SKU sales in all locations. This sample size showed interesting results within satisfactory running time for the analyzed set. Training on larger sample sizes leads to exponentially increasing run time. Therefore, the completeness of data must be balanced with the computational costs while producing accurate classification of SLC and LLC items in a representative sample.

Granularity level: The default granularity of the framework is at SKU level in all locations. Still, users can provide specific escalations for analysis if it is in accordance with the uploaded hierarchies.

Outlier threshold: Data preprocessing requires treating extreme values caused by stock-outs, promotions, or outliers. This procedure applies the calculation of the deviation range from the trend in the observed series. Punctual sales above or below specific limits are treated as extreme values. The threshold is obtained as the number of standard deviations above or below the smoothed trend. This is a user defined parameter that models the sensitivity to outliers and assumes a default value of 3.

Smoothing filter & **Parameter**: Noise reduction is an integral part of the proposed solution. Moderates the randomness of data and the impact of large variations. The model accepts the smoothing of data using the Moving Average (MA) or Gaussian Filter (GF). These approaches have specific characteristics and effects on data. The model can be selected by the user as (‘Gaussian’ or ‘MA’) and the respective parameters required, n and σ, must be provided.

Cluster Model: The clustering techniques supported by the framework are k-means and DBSCAN. These methods have different clustering approaches and fit data characteristics. The parameters for k-means and DBSCAN, *k*, *eps* and *MinPoints*, respectively, are automatically optimized for the model by evaluating the silhouette score indicator, not requiring user input. Even though the model defaults to the DBSCAN algorithm, the clustering method can be specified by the user.

Forecast window: When it comes to the forecast analysis, demand prediction can be evaluated under several conditions. The number of future periods to be foreseen impacts the forecasting error. Predicting further into the future will usually lead to lower accuracy when compared to short-term forecasting. Setting the forecast window depends on the business requirements and reliability of the data. For the analyzed samples, forecasting was performed over the short and medium term, 1 and 8 weeks (about 2 months) into the future, respectively.

Test window: The size of the test window controls the representativeness of the obtained forecast metrics. Testing on sizable data leads to the understanding of the forecasting model performance over a longer period, improving the reliability of the results. On the other hand, too large of a test set may lead to reduced data for the adequate training and validation of the model. The test window is configurable, and 16 weeks (about 3 and a half months) was considered for the default analysis.

Number of folds: Cross-validation is a technique employed to evaluate a model in various sections of the data. Cross-validation for sequential data produces subsequent sections of the time series to train the model with unseen data and evaluate the predictions on each fold. The application of this technique was twofold in the forecast component. For parameter optimization (validation set) and evaluating performance (test set). The default number of folds is 3.

4.3 Data preprocessing

The data preprocessing methodology is addressed to prepare data for posterior analysis. The main goal of this section is to present an in-depth description of the procedures adopted, referring to the resources employed, and the assumptions stated.

To identify general patterns that may characterize products demand as SLC and LLC, it becomes necessary to treat sales data in a way that the demand behaviors for these product types are highlighted and distinguishable. This process deeply impacts the representation of the data and the conclusions drawn by the analysis. The procedure involves considering techniques to prepare data over the main preprocessing concepts: granularity, extreme values, noise, and scale. The result at the end of this stage is a set of products' demand time series with lower dimensionality than the original set. The produced set is posteriorly used for classification and forecasting purposes.

It is relevant noticing that, for forecasting, each time series was considered ranging within the first and last sales registered for the respective item. Whereas for clustering, the time series were ranged within the whole 3 years collected, filling absent sales with 0.

4.3.1 Aggregation

The degree of aggregation has a significant impact in the classification and prediction of the products' demand. In a general line, lower levels of granularity are characterized by reduced specification of sales and represent a higher-level overview of the products demand. Finer granularities on the other side represent more specific sales contexts but may register scarcer data.

Escalation (aggregation) of sales data is conducted under defined hierarchy structures. Hierarchies map the sales degree of specification where higher hierarchy levels are associated with larger sized and less specific groups. Hierarchies are sub divided into two major structures: product, and location. The tables presented in *Fig*. 4, list the various escalation levels structured for the product and location hierarchies accepted in the RDF tool.

When it comes to the product hierarchy, the standard base granularity in the retail industry is typically the Stock Keeping Unit (SKU). This level of aggregation is the most specific and details the color, size, or other item attributes. On the location hierarchy, the store is usually at the base escalation level.

The forecasting of sales data is required at the base units of product and location to favorably estimate the exact SKU quantities to be allocated to each store. When the presence of historical data at the finest escalation level is not significative, aggregation is a widely applied procedure cross-industry. With this process, forecasts are produced at lower levels of aggregation and then estimated for the higher level using adequate methods.

Name	Label
SKU	Item
SKUP	Style/Color
SKUG	Style
SCLS	Sub-Category
CLSS	Category
DEPT	Department
PGRP	Group
DVSN	Division
CMPP	Company
VNDR	Vendor
PAT1	Prod Attribute 1
PAT2	Prod Attribute 2
STA1	Style UDA 1

Name	Label
STOR	Location
DSTR	District
REGN	Region
CHNL	Channel
CHAN	Chain
COMP	Company
SFMT	Store Format
STCL	Store Class
SAT1	Store Attribute 1
SAT2	Store Attribute 2

Fig. 4 –Product (left) and Location (right) standard hierarchies in RDF.

In the problem's context, it is intended to study the demand patterns that characterize products as SLC and LLC at the base product level (SKU). The tests were conducted at SKU product level and company (COMP) location level. In this analysis the location factors were disregarded and considered solely factors related to the product.

4.3.2 Extreme values treatment

As previously stated, extreme values are accountable to stock-outs, promotions, or outliers in the retail industry. Forecasts are not reliable when the algorithms are trained with these observations, statistical methods assume stable patterns that can be generalized to predict future demand.

Extreme variations are hardly predictable and introduce random complexity to the analysis. In these observations, algorithms may under or over-fitting data or not converge at all. This concludes about the relevance of treating these types of data event, becoming critical to promptly identify and correct them. One can approach this process by defining extreme values as observations that deviate a certain number of standard deviations above the mean of the time series data. This way, the defining threshold is set uniformly for all data points.

However, a more elegant procedure is considered to identify the extreme values threshold range as a defined number of standard deviations away from the trend component of the time series. This generates a dynamic threshold range that accompanies the smoothed trend. To isolate the trend component from the original time series, the model applies a Gaussian Filter with parameter $\sigma = 4$. Then the environmental variable, *Number of Standard deviations*, is multiplied by the standard deviation and added or subtracted to the trend at each instance to produce the normal values acceptance range. The plot in *Fig*. 5 presents the extreme value detection procedure with the sales of a single SKU sampled from the complete set.

After identifying the extreme observations, it is necessary to correct these values. This is a tricky process since the relevance of the extreme observations varies for each of these event types. Promotion and stock-out events happen in a planned or random manner; therefore, these instances add little to no effect on the expected sales variation and are not predictable of future demand. Outliers on the other side may happen due to a variety of external factors and may be interpreted as the extrapolation of the expected deviations. In the industry scenario, internal events are registered and used to correct the extreme values allegedly caused.

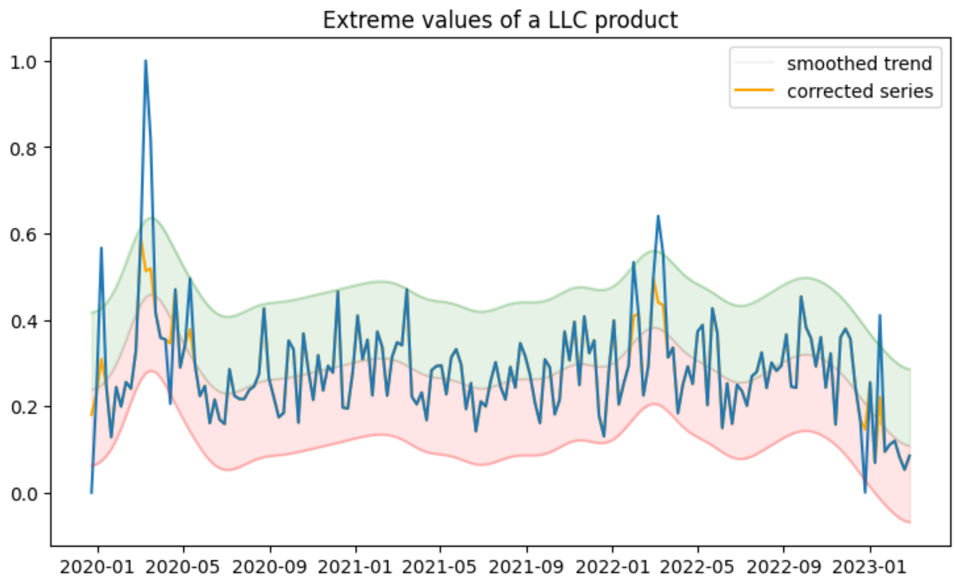

Fig. 5 – Outlier detection for an LLC product time series.

To adequately smooth the extreme values, promotions and stock-outs should be corrected with the average between the previous and subsequent sales. Outliers, however, can be interpreted as variation in the data and corrected with the standard deviation above or below the trend component. During the project, promotion and stock-out data was inaccessible. With this limitation, the procedure adopted to correct extreme values considered replacing them by normal deviations from the trend, as shown in *Fig*. 5.

4.3.3 Noise smoothing

The presence of noise, i.e., unexplained variation in the data, may obscure the significant patterns and make them difficult to identify.

This presents a challenge for some machine learning methods which are sensible to this phenomenon, such as the k-means and DBSCAN algorithms, as well as the statistical forecasting methods considered. For the clustering algorithms, the noise level affects the selection of the optimal parameter and consequently the quality of the labels generated. In the presence of noise, these methods may confuse random variation as patterns or vice versa, which may lead to inaccurate distance measure computation and inadequate clustering.

With this scenario, noise smoothing is mandatory, and this procedure was implemented with two alternative smoothing methods, **Gaussian Filter** (GF) and **Moving Averages** (MA).

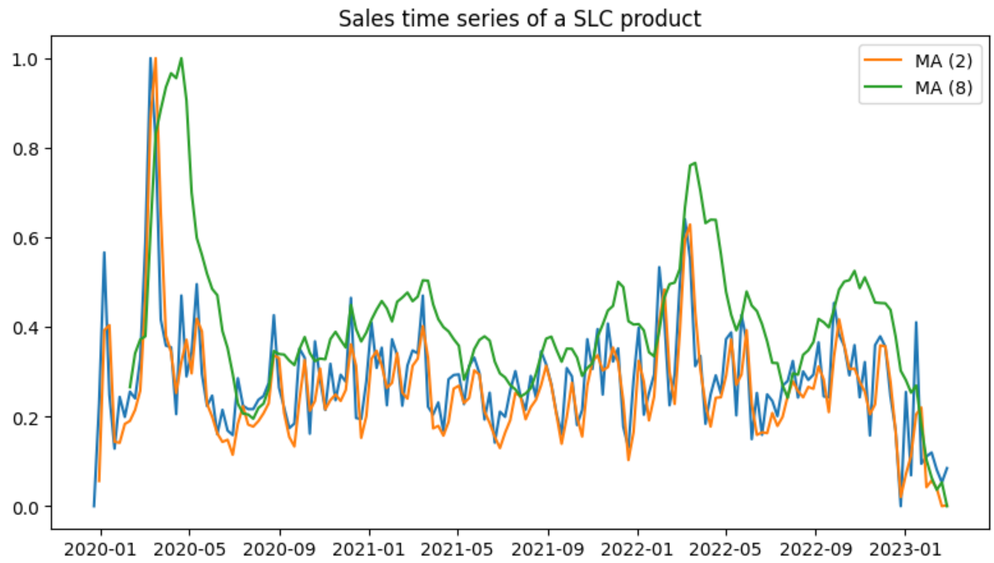

Fig. 6 – MA smoothing for an LLC product time series ($n = 2$ and $n = 8$).

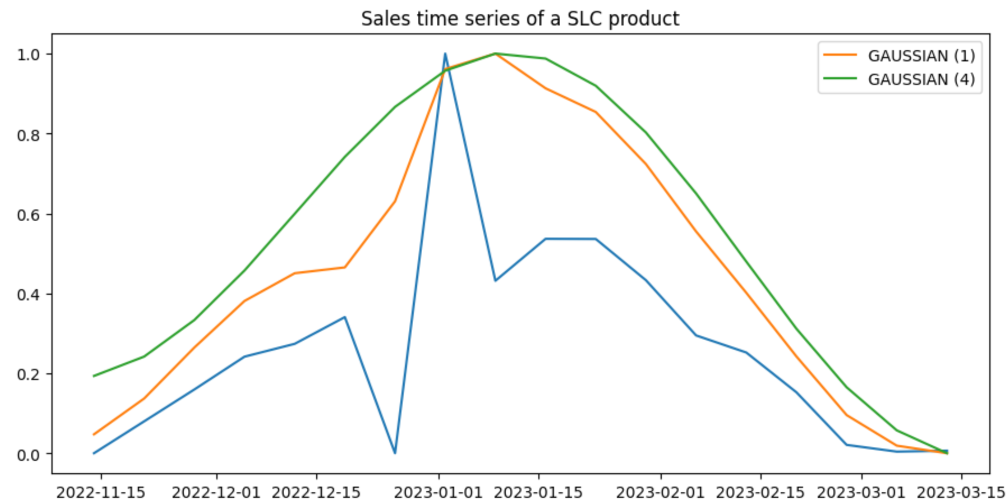

Fig. 7 – Gaussian smoothing for a SLC product time series ($\sigma = 1$ *and* $\sigma = 4$).

Both the approached methods are suitable to smooth time series data. However, the gaussian filter properties made it a suitable candidate to preprocess data prior to clustering. On the other hand, and due to the common application of moving averages to preprocess data for forecasting, this method was selected for this end.

4.3.4 Normalization

The last stage of the proposed data treatment process is normalization. As introduced in Chapter 3, the time series were individually normalized, on a scale ranging in [0, 1], so that 0 represents the actual absence of sales and 1 the sales peak of each time series.

The absolute 0 of sales data represents the absence of sales. When normalizing this data, considering the representation of the absolute 0 leads to normalized series where the variation is relativized according to the mean sales level. This way, even though the sales volumes are not considered, the relative levels of sales remain intact. Meaning that the relative level where variation revolves is considered and comparable across the data set, regardless of the actual sales volume.

To achieve this effect, the framework bases the normalization procedure on the Min-Max technique described in chapter 2. The lower boundary (*min(x))* value was set as 0, resulting in the simplified *Equation* 25.

$$x_{scaled} = \frac{x}{max(X)} \qquad \textit{Equation } (25)$$

This normalization method produces time series scaled in [0, 1], with 0 as an absolute value that represents absent sales.

4.4 Clustering of time series

In this Section, the implemented clustering methods for time series are described in detail. These are unsupervised methods that explore relationships in the data without requiring a known target variable for training.

The clustering techniques considered were referenced in Chapter 2: k-means and DBSCAN. These are popular clustering algorithms and can be used for time series data by computing the distance between series.

The data space was generated based on the DTW distances between series. This metric was used to represent time series as data objects to be clustered. The DTW procedure is exemplified in *Fig*. 8. The distance between two unsynchronized time series is calculated in the optimal alignment.

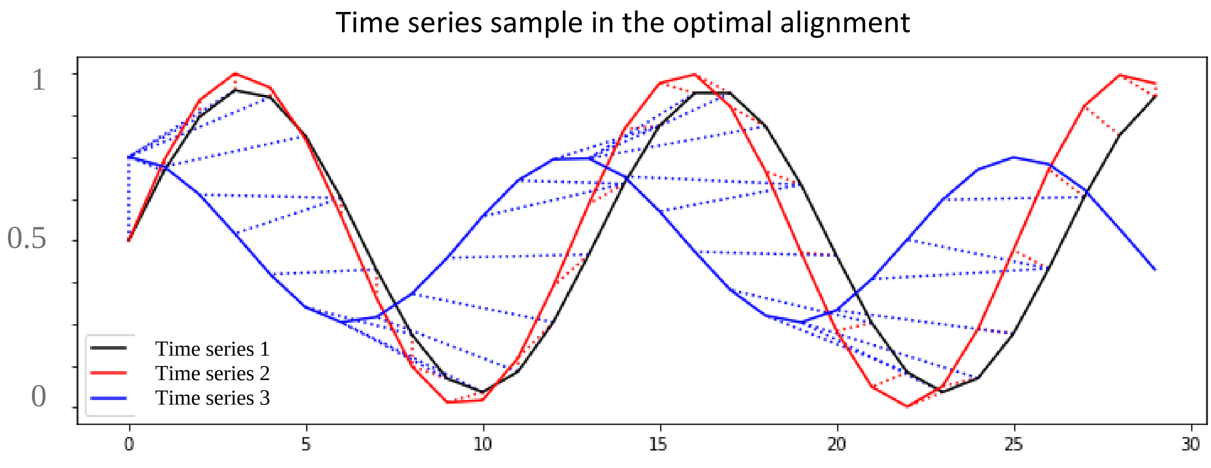

Fig. 8 – DTW approximation between two time series.

The graph of *Fig*. 9 shows a 2-dimensional multidimensional scaling representation of the sample analyzed. The DTW distances were considered to produce this representation and even though the axes of the plot do not have any inherent meaning, they reflect the relative distances between the points. The plot defines a clear dense region with a linear shape, while the points outside of this region are dispersed and do not show clear nor coherent relationships. The hypothesis is that the identified dense region groups the SLC items and the disperse data points are related to LLC products with particular demand behaviors. The improvement of the sample analyzed, both by increasing the sample size and stratification of the complete data, may consolidate patterns in other regions of the data space, further developing the proposed framework.

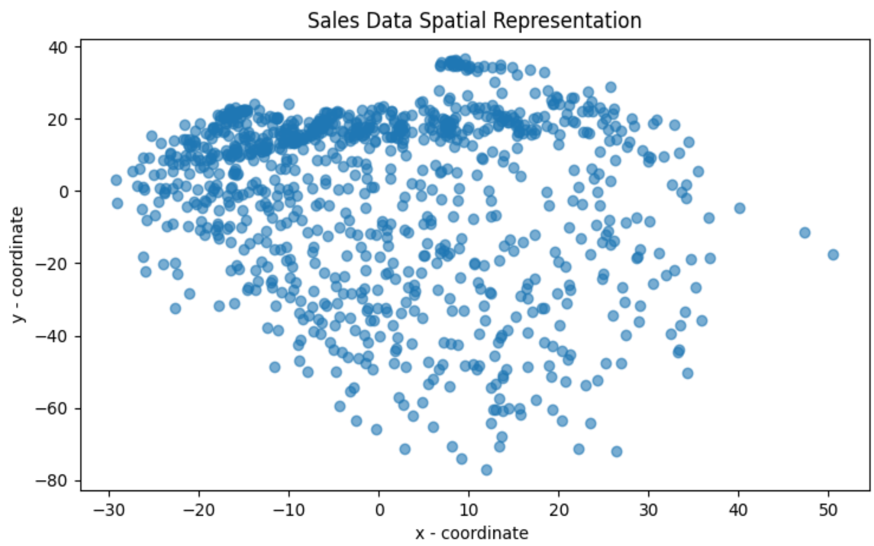

Fig. 9 – Spatial representation of the sample in a 2-dimensional space.

This Section is intended to describe the clustering approach of the proposed methodology, presenting the two optional clustering approaches to the problem in study.

4.4.1 K-means

The k-means algorithm is a partition-based clustering technique that was implemented to identify SLC and LLC products from time series data. The optimal parameter *k* was selected by minimizing the silhouette score.

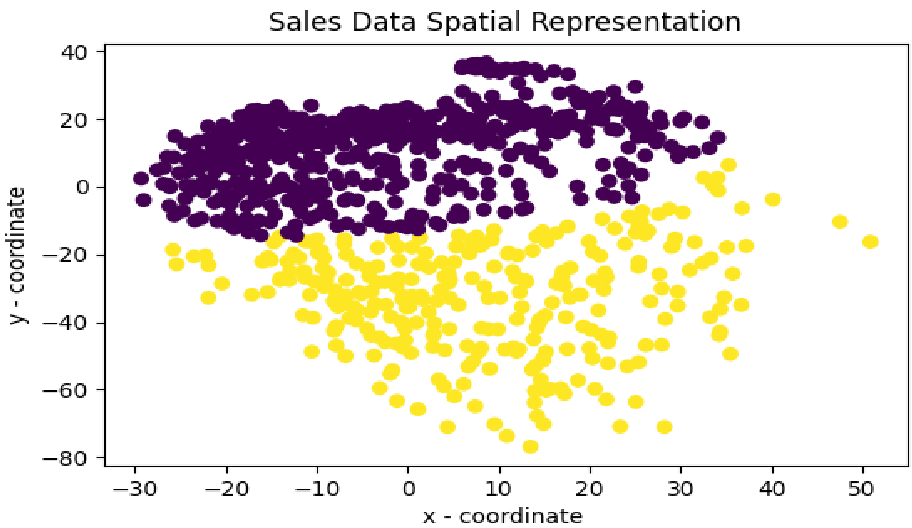

Fig. 10 – Spatial representation of the sample, colored by k-means labels.

This approach was shown to be simple yet effective for grouping data into homogeneous partitions when the number of clusters, *k*, can be determined a priori. The graph in *Fig*. 10 plots the optimal *k* clusters generated by k-means with the products sampled from the complete data set using stratified sampling, in a two-dimensional spatial representation.

4.4.2 DBSCAN

DBSCAN is a density-based algorithm that discovers clusters of arbitrary shape. The method can handle data with varying density and adapt to its underlying structure. This technique was implemented to identify SLC and LLC products from time series data. The optimal parameters *eps* and *MinPoints* were selected by silhouette score minimization.

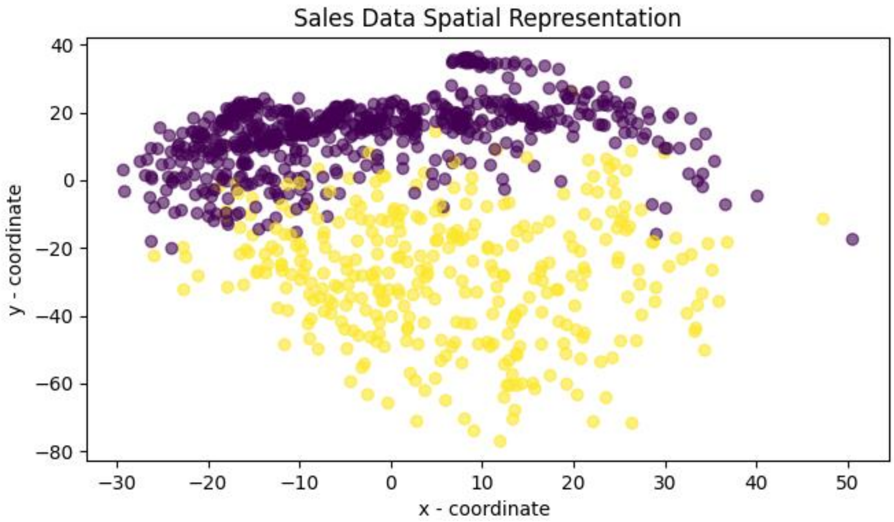

Fig. 11 – Spatial representation of the sample, colored by DBSCAN labels of dense clusters.

The analysis of the spatial representation of the data points clustered with DBSCAN *Fig*. 11 indicates the existence of a dense region of linear shape labeled as SLC (purple) and a sparser conglomerate labeled as LLC (yellow). By comparison of the representations in *Fig*. 10 and *Fig*. 11, it is conclusive that k-means produced homogeneous partitions of the space whereas DBSCAN clustered areas of dense points.

4.4.3 Parameter optimization

The clustering methods proposed require the definition of parameters. K-means using the number of *k* clusters to group. DBSCAN on the other hand models two parameters: neighborhood radius (*eps*) and minimum neighbor points (*MinPoints*).

Finding the best fitting parameters for the sample in analysis is performed by optimizing the silhouette score for all parameter combinations. Parameter tuning refers to the procedure of choosing the optimal parameters. This metric is useful to evaluate the performance of clustering models and can be used to guide the selection of the parameter *k* in the k-means algorithm or the *eps* and *MinPoints* of DBSCAN. For that purpose, the overall Silhouette Score was calculated for the models fitted with each parameter combination, and the best performing parameters were selected, *Fig*. 12.

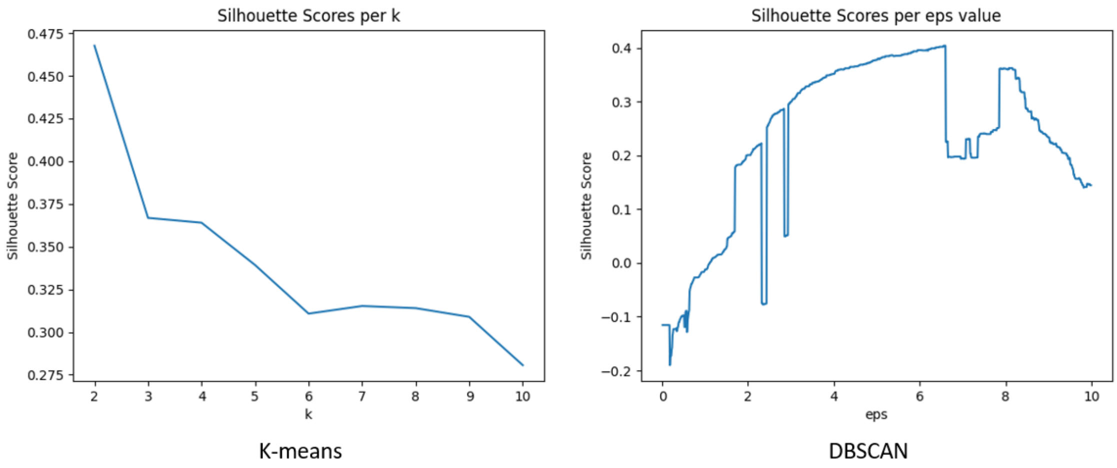

Fig. 12 – Plot of the silhouette scores for k-means and DBSCAN with the tested parameters.

In summary, the best fitting parameters for both k-means and DBSCAN resulted in the formation of two clusters, that are expected to classify the demand behaviors of the SLC and LLC products.

4.5 Demand forecasting

Demand forecasting is a widely researched field of study. Accurate forecasting is a crucial aspect for the retail context. Therefore, a growing number of techniques have been proposed for the prediction of future demand and adapting to data qualities.

Some of the most widely used forecasting techniques are the Simple Exponential Smoothing (SES), Holt-Winters (HW) and the Auto Regressive Integrated Moving Averages (ARIMA). These models were selected to evaluate the generated classes due to their distinct characteristics that allow for different evaluation approaches.

While SES performs well with simple time series, HW is effective in the presence of predictable trend and seasonal components. ARIMA finds autocorrelation in the series and is a more flexible model that can fit complex dynamics. Since SLC products may lack sufficient data and consolidated patterns, it is expected that LLC items show better overall forecast performance than short life cycle ones. The advantage of using the selected methods in detriment of more complex ones that better suit SLC items lies on their simplicity and low computational cost.

4.5.1 Cross-validation and parameter tuning

Cross-validation is a technique used to assess the performance of a given model across different partitions of the train and test data. Since temporal sequences lose meaning if randomized, cross-validation can be applied to time series by partitioning the test in subsequent folds.

The defined model is used to predict each of these folds, and the error metrics are computed. The final performance is obtained by the average forecasting metrics measured over the *k* folds. This method leads to an overall perspective of the model performance in different sections of the evaluation set by providing a more robust measure of error.

For that reason, this technique was used for both parameter optimization and performance evaluation purposes. The adopted strategy for parameter optimization consisted of subsequently training and testing in different sections of the train and test sets. The forecasting component of the framework aims to evaluate the predictive performance of each escalation. Users can define the forecast window and the test window (weeks reserved for testing). An additional parameter is featured to select the k partitions of the validation and test windows.

For the analysis of the data set, forecasting was performed over test windows of 16 weeks with forecast window of 1 or 8 weeks for the short and medium terms, respectively. The framework assumes 3 folds for cross-validation, as illustrated in *Fig* 13.

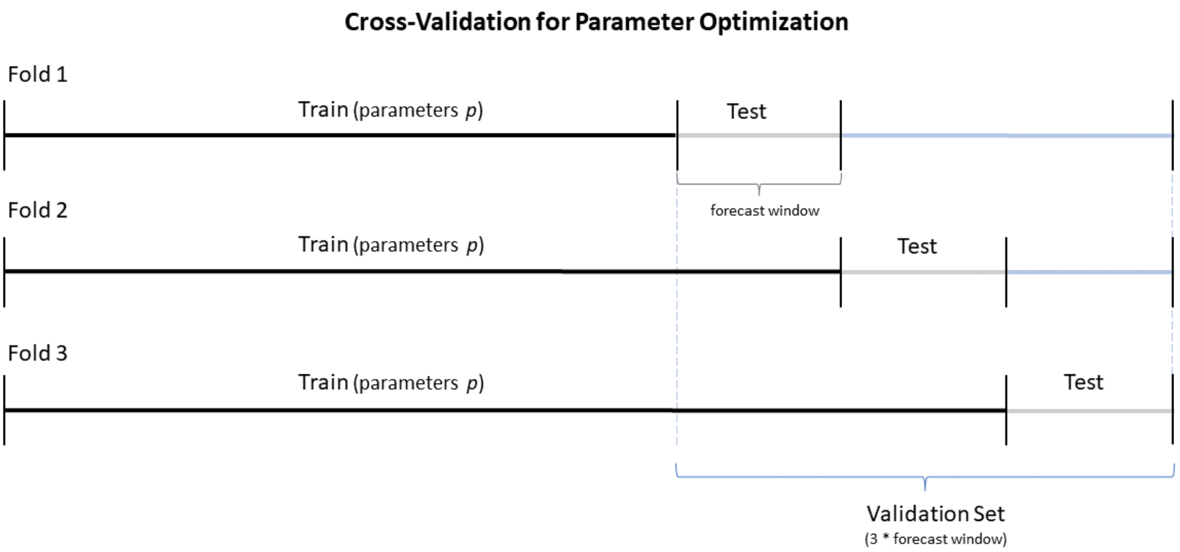

Fig. 13 – Cross-Validation process for parameter optimization on a validation set

For forecasting parameter tuning, cross-validation was used to select the parameters that lead to the best fitting model for each item. The implemented forecasting models (SES, HW, ARIMA) require the definition of parameters. Selecting the best values as parameters profoundly impacts performance. Therefore, they must be estimated to ensure the reliability of the forecasts.

The adopted framework cross-validates a wide range of feasible parameters in the validation set. The exponential smoothing methods, SES and HW, are fitted with the possible combinations of $\alpha \in [0.1, 0.2, 0.4, 0.6, 0.8]$, $\beta \in [0.1, 0.3, 0.5, 0.7]$, and $\gamma \in [0.1, 0.3, 0.5, 0.7]$. Where β and γ are exclusively required for HW. When it comes to ARIMA model selection, the ***auto_arima*** function from the *pmdarima* package combines stepwise and information criteria to determine the best parameters automatically. The *auto_arima* function uses the log-likelihood *Equation* 26 to calculate the Akaike Information Criterion (AIC), Hu, S. (2007), Equation 27, and the Bayesian Information Criterion (BIC), Bhat, H. S., & Kumar, N. (2010), *Equation* 28.

$$L_k = -\frac{n}{2} * \ln(2\pi) - \frac{n}{2} * l\,n(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{n} (x_i - \mu)^2 \qquad Equation\ (26)$$

$$AIC_k = -2 * \ln(L_k) + 2 * k \qquad Equation\ (27)$$

$$BIC_k = -2 * L_k + \ln(n) * k \qquad Equation\ (28)$$

Where k is the number of independent variables to be estimated, L_k the log-likelihood, and n is the number of observations.

This method allows the automation of parameter selection for large data. Even though it reduces the level of expertise required to produce quality forecasts, it may converge to sub optimal solutions, and require finer analysis.

4.6 Framework assessment

4.6.1 Test design

This Section explores the test design methodology implemented to access the degree of achievement of the proposed goal of identifying SLC products and separating them from LLC. The tests were conducted by defining hyper-parameters and running the customized framework. The assessment then takes place, and the indicators are interpreted.

The evaluation of the proposed framework is undertaken over two dimensions: **clustering performance** and **capability of classifying** products as SLC and LLC. The clustering performance is assessed through inspection of the spatial relationships and interpretation of the silhouette scores. The classification accuracy is empirically quantified through the analysis of the forecasting performance of the three methods previously mentioned, combined with indicators associated with each of the product types.

The first *test* considered a sample of size $N = 1000$, randomly selected from the original set (**simple random sampling**). Another sample of equal size was taken for *test* 2 and *test* 3, this time by randomly selecting from groups in the data (**stratified sampling**). The set was stratified by the time interval elapsed between the first and last sales (selling period). Products with over 2.5 years of registered sales were grouped in the first stratum and the remaining in the other. The sample procedure selected 500 products from each of the produced strata to constitute the sample.

In *test* 1 and *test* 2, the *k* clusters were created using the **k-means** algorithm. The clusters created by this method are homogeneous and the density of data points is not directly considered. Further in *test* 3, the same sample of *test* 2 is labeled using the **DBSCAN** algorithm. The method created clusters that consider density regions in the data and assume any density-based pattern.

The forecasting procedure proposed is applied across all tests. SES, HW, and ARIMA are used to forecast the demand of each product in the sample. For each test, the forecast was performed over short-term and medium-term objectives. Setting the forecast window to 1 week (short-term) and 8 weeks (medium-term). The evaluation of the forecast performance was estimated over a 16-week window for all tests.

Tests design to evaluate the framework proposed

	Test 1		Test 2		Test 3	
	a) Short	b) Medium	a) Short	b) Medium	a) Short	b) Medium
Sample method	SRS	SRS	SS	SS	SS	SS
Sample size	1000	1000	1000	1000	1000	1000
Cluster method	K-means	K-means	K-means	K-means	DBSCAN	DBSCAN
Forecast window	1	8	1	8	1	8
Test window	16	16	16	16	16	16
k folds	3	3	3	3	3	3

SRS - Simple Random Sample
SS - Stratified Sample

Fig. 14 – Proposed hyper-parameters for evaluation of the framework.

4.6.2 Empirical approach

If the characteristics widely associated to SLC and LLC products are assumed linearly relatable, an empirical model can be developed to identify SLC and LLC products. For interpreting this approach's results, accuracy metrics are constructed to evaluate the proposed solution.

In the absence of labeled data, a simplistic empirical model that classifies SLC as the linear combination of the proposed indicators based on threshold values can be constructed, *Equation* 29. This method is a simple representation of the relationships between SLC and the dependent variables and, even though no accuracy measure is quantified across actual data, it can be employed to further study the adequacy of the proposed framework. This approach relates binary indicators by assuming thresholds and do not directly explore patterns in the demand data. Therefore, the proposed clustering approach may lead to a further refined classification.

$$LC_t = \frac{1}{n}(x_1 + x_2 + x_3 + x_4 + \cdots + x_n) \quad Equation\ (29)$$

Where *LC* is the linear combination of $x_{i\ \epsilon\ (1,2,..5)}$, the binary variables of *CV*, *P*, *Nsales*, *SMAPE*, and *ME*, obtained by segmenting the indicators above or below the predetermined thresholds *Equation* 29.

The linear combination can be further developed by adding weights to the binary variables according to the relative importance of the indicators in the classification of SLC and LLC items, *Equation* 30.

$$WLC_t = \frac{1}{\sum_{i=1}^{n} w_i}(x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + x_4 * w_4 + \cdots + x_n * w_n)$$
$$Equation\ (30)$$

When combined with the proposed classification, the weighted average of life cycle binary indicators is a simple yet effective method to classify SLC and LLC items based on accessible metrics. The labels obtained by time series clustering are used to optimize the indicators weights. This model can then be applied to classify products automatically based on machine learning knowledge and without requiring further train data.

To estimate the thresholds used to identify the SLC, it is important to consider the characteristics referenced for these products. Typically, SLC items are marked by soaring coefficients of variation and intermittency due to the irregularity of the data. SLC are also associated with short times in the market (from a few weeks to a few years). These factors usually lead to low accuracy and biased forecasts for these products, when using statistical forecasting methods like SES, HW, or ARIMA.

Thresholds were defined for each of these features to transform them into binary variables. To qualify high relative variation in the demand, the coefficient of variation (***CV***) threshold was set as 1 (where standard deviation is equivalent to the mean sales). For intermittency (***P***), the reference value 1,32 was used to classify intermittent products. The maximum number of sales (***NSales***) selected was equivalent to two years of data, 104 weeks. The ***SMAPE*** threshold selected was 30%, representing poor forecasting accuracy. SLC products are also associated to negatively biased forecasts with traditional methods, so the ***ME*** threshold was set to 0.

5 Results

Over this Chapter, the results of the conducted tests, referenced in Chapter 4, are thoroughly presented. Section 5.1 approaches the procedure that correlates the clusters to characteristics frequently associated to SLC and LLC products. In Section 5.2, the indicators resulting from the forecast experiments are demonstrated. Section 5.3 presents the results of correlation with the indicators, coefficient of variation, number of weeks with sales, intermittence, and metrics, *SMAPE*, *ME*.

5.1 Classification

Under the evaluation of the classification methodology, the traits of the classified products are registered and correlated, supported by visual analysis of the time series.

The samples of the identified groups express visual differences and similarities among product types. SLC samples show **abrupt trend shifts** over **short periods**.

Visually, SLC items show a characteristic triangular shape marked by the prevalence of **abrupt growth** followed by **steep decline** and **transient maturity**. On the other hand, LLC registers long sales periods marked by demand components like seasonality, level, or trend, and expresses stable patterns over extended periods, in *Fig*. 15.

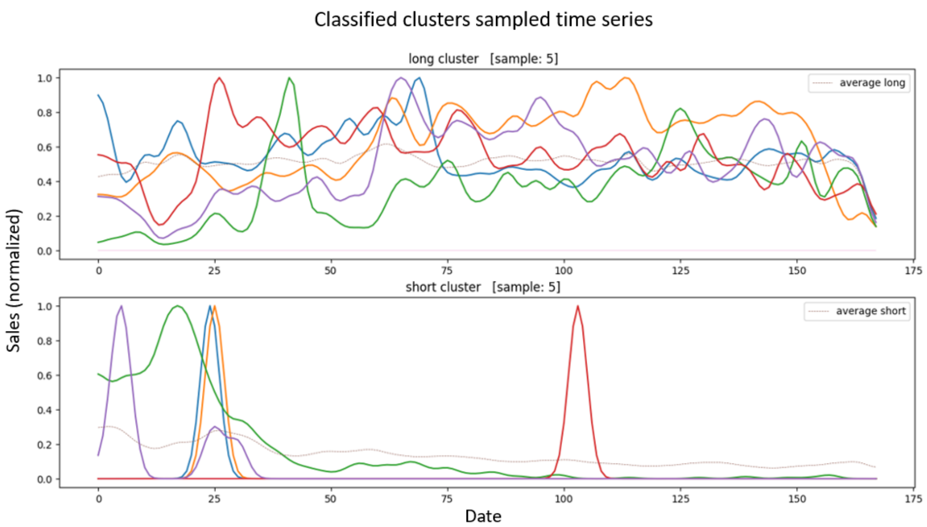

Fig. 15 – Product time series classified in *test* 3 with DBSCAN.

The visualization of samples like the one in *Fig*. 15, even though informative, is not quantifiable. To analytically interpret the cluster classification for each of the clustering methods, an extensive analysis was conducted. Three **demand indicators** (coefficient of variation, number of weeks with sales, and intermittency), and two **business indicators** (sales volume, and sales period) were considered to evaluate the proposed classification.

5.1.1 Coefficient of variation

With both clustering approaches, the coefficient of variation of the SLC and LLC classified groups are differently distributed. For both methods, **SLC** expressed significantly **higher levels** of relative variation when compared to LLC.

Table 1 – Coefficient of variation: Q_1, median and Q_3, by test and class.

Test	method	class	Q_1	median	Q_3
Test 2	DBSCAN	SLC	0.756	1.000	1.323
		LLC	0.519	0.723	1.070
Test 3	k-means	SLC	0.781	1.028	1.356
		LLC	0.492	0.646	0.920

5.1.2 Number of weeks with sales

The number of weeks with sales was shown as a **decisive** indicator of the SLC and LLC groups generated with both tests. SLC clearly presents a lower number of sales than LLC items. The average **SLC** group had ¾ of the items with **less than 87 weeks** (about 1 year 8 months) of sales. On the other hand, ¾ of **LLC** items had **more than 164 weeks** (about 3 years) of sales.

Table 2 – Number of weeks with sales, Q_1, median and Q_3, by test and class.

Test	method	class	Q_1	median	Q_3
Test 2	DBSCAN	SLC	14	34	80
		LLC	163	165	165
Test 3	k-means	SLC	16	36	93
		LLC	165	165	165

5.1.3 Intermittency

Intermittency also marked disperse differences among the SLC and LLC classes. The **SLC** group *p* (intermittence) dispersed over sizable values **above 1.18**. Whereas **LLC** products *p* values displayed a tight distribution **centered** in ~**1**.

Table 3 – Intermittency: Q_1, median and Q_3, by test and class.

Test	method	class	Q_1	median	Q_3
Test 2	DBSCAN	SLC	1.024	1.174	1.929
		LLC	1.006	1.006	1.006
Test 3	k-means	SLC	1.025	1.179	1.987
		LLC	1.006	1.006	1.011

Intermittency values with $p > 1{,}32$ are considered intermittent products in the sector. The SLC group contained 40.69% and 41.24% intermittent items for the k-means and DBSCAN approaches, respectively. Whereas the LLC groups contained 1.43% and 6.55% intermittent products.

Across the results, the k-means methodology of *test* 2 produced more concise LLC groups according to coefficient of variation, number of weeks with sales and intermittent, when compared to DBSCAN of *test* 3. The results of the correlation between the demand indicators and the grouped items are summarized in *Fig.* 16.

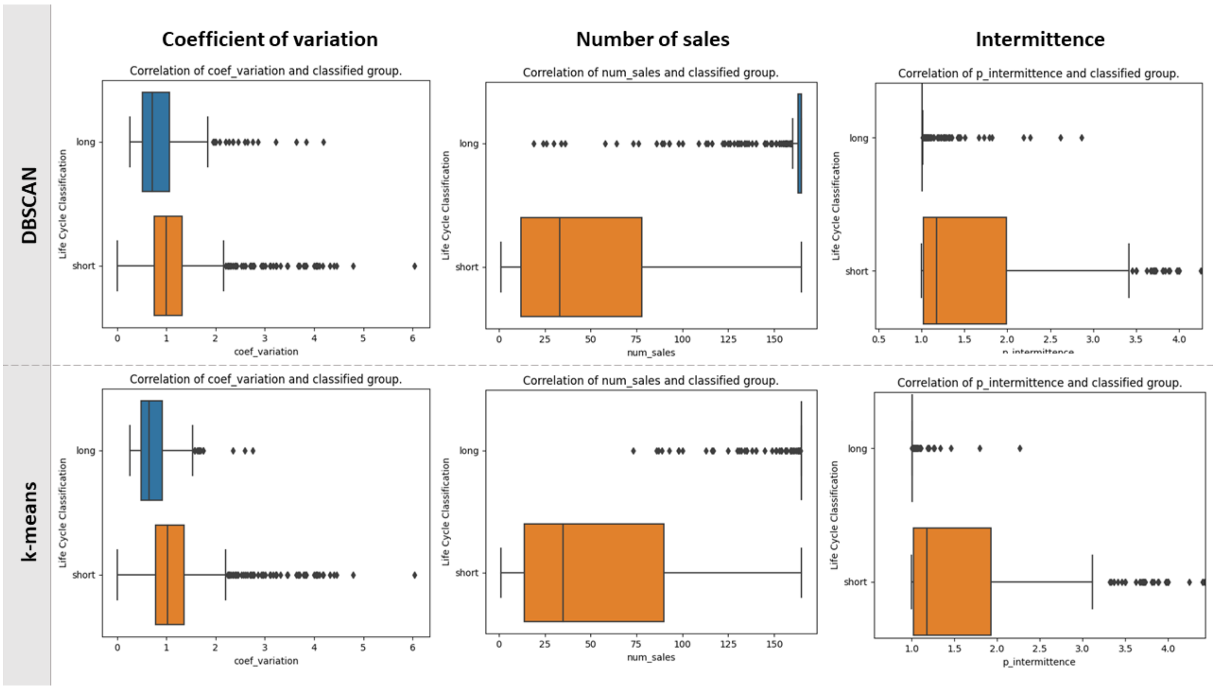

Fig. 16 – Correlation summary of the demand indicators per: SLC (orange) and LLC (blue).

5.1.4 Sales volume

The average sales per year marginally differed between the SLC and LLC groups, where **SLC** were associated with **lower** and **LLC** with **larger** sales volumes. In the DBSCAN approach, the groups comprehend similar distributions, not being considered significative to clearly difference the classes.

Table 4 – Sales volume (units), Q_1, median and Q_3, by test and class.

Test	method	class	Q_1	median	Q_3
Test 2	DBSCAN	SLC	266	4231	35705
		LLC	3511	15204	63649
Test 3	k-means	SLC	251	5429	39226
		LLC	2438	11977	49948

5.1.5 Sales period

The number of weeks elapsed from the first and last registered sales is a business indicator that can characterize the groups. ½ of the **SLC** group register sales range **lower than 50 weeks** (about 11 and a half months), the equivalent value for the **LLC** was **166 weeks** (about 3 years).

Table 5 – Sales period (days): Q_1, median and Q_3, by test and class.

Test	method	class	Q_1	median	Q_3
Test 2	DBSCAN	SLC	28	45	154
		LLC	166	166	166
Test 3	k-means	SLC	30	56	159
		LLC	166	166	166

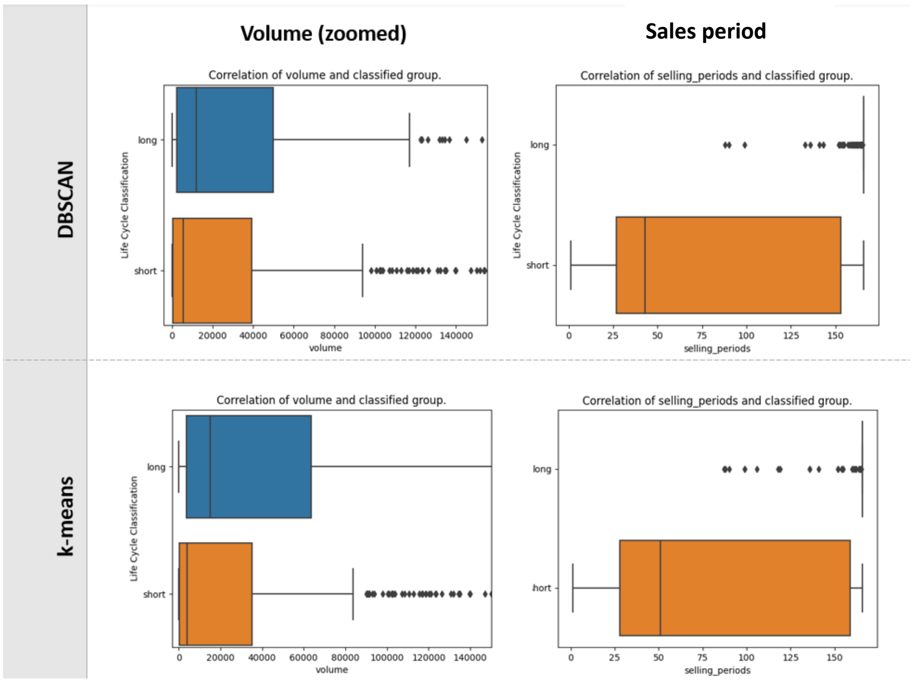

Fig. 17 – Correlation summary of the **business indicators** per: SLC (orange) and LLC (blue).

5.2 Demand forecasting

To evaluate the framework under the forecasting perspective, this section summarizes the performance metrics results for the SES, HW, and ARIMA procedures. The results are presented in the light of the previous Section, where the dispersion of each calculated metric is characterized within the groups.

Due to the similarity of drawn conclusions between the tests conducted and for the simplicity of the report, only the results from *test* 3 for medium-term forecasting are detailed in this section. The results for this component are focused on two of the performance metrics calculated, SMAPE and ME.

5.2.1 SMAPE

The SMAPE results expressed a clear distinction between distributions of SLC and LLC classified products across all forecasting methods. When it comes to **SLC** forecasting, SMAPE is usually extreme, and, on average, ¾ of SMAPE values were **above 48**%. **LLC** express modest variations of SMAPE, and for the 3 models, ¾ of these items achieved values **below 40**%. These results show high separation between classes and suggest that the SLC class experiences high forecasting error when forecasting on medium-term whereas LLC presents more moderate values, *Fig*. 18.

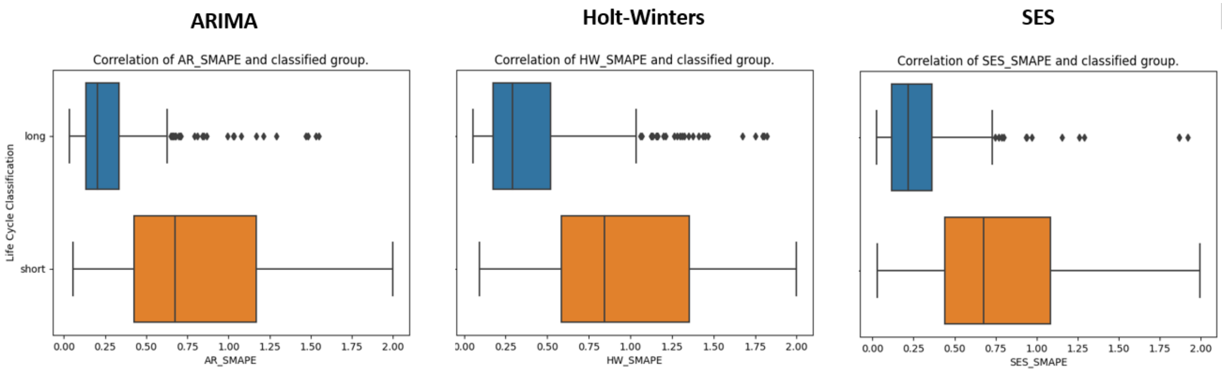

Fig. 18 – SMAPE results across SLC (orange) and LLC (blue), by forecasting model.

Comparing the performance of the forecasting models, the ARIMA was the best median performer for SLC and LLC, with median values of 0.675 and 0.204, respectively. Followed by SES and the worst performer Holt-winters.

Table 6 – SMAPE: Q_1, median and Q_3, by method and class.

method	class	Q_1	median	Q_3
ARIMA	SLC	0.428	0.675	1.167
	LLC	0.135	0.204	0.336
Holt-Winters	SLC	0.586	0.844	1.335
	LLC	0.176	0.292	0.521
SES	SLC	0.438	0.676	1.083
	LLC	0.118	0.215	0.363

The poor performance of the HW when compared to SES and ARIMA may be explained by incomplete parameter optimization. This may lead to under or over-fitting and consequently aggravates the forecast error.

5.2.2 ME

Following with the results of the mean errors, both groups are centered near 0. However, the dispersion of ME results differs from SLC to LLC. The group of **SLC** show prevalent negative bias, where ¾ of the ME results were **below 0.037**, on average. The **LLC** class on the other hand suggests ME values normally distributed and **centered in 0**, across the forecasting methods. The performance measured with ME shows that Holt-Winters and SES were skewed whereas ARIMA distribution was nearly centered for the SLC class, *Fig*. 19.

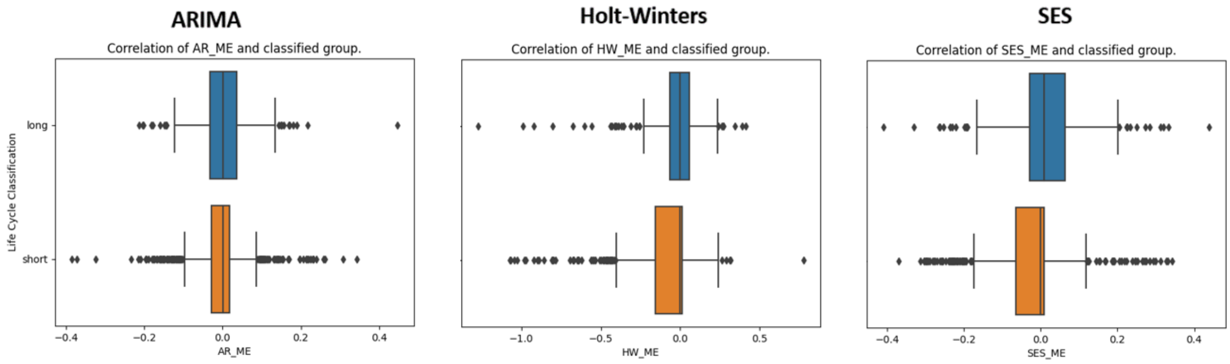

Fig. 19 – ME results across SLC (orange) and LLC (blue), by forecasting model.

Table 7 – ME: Q_1, median and Q_3, by method and class.

method	class	Q_1	median	Q_3
ARIMA	SLC	-0.028	0.000	0.017
	LLC	-0.033	0.001	0.036
Holt-Winters	SLC	-0.154	-0.001	0.011
	LLC	-0.067	0.000	0.055
SES	SLC	-0.064	0.000	0.009
	LLC	-0.029	0.008	0.064

5.3 Comparison to linear combination

The current Section presents the evaluation of labels produced with the framework against the linear combination (empirical model) described in Section 4.6.

This analysis estimates the agreement between the clusters generated and the empirical linear combination with thresholds simplistic model. Both k-means and DBSCAN approaches performed accurately when compared to the linear model.

DBSCAN led to improved precision and specificity with values 87.3% and 84.1% in contrast with 86.8% and 82.3% of k-means. The latter, however, surpassed DBSCAN in the recall value with 83.1% and marginally in the accuracy and f1 score values.

Overall, the framework was in accordance with the simplistic model, quantified by the performance indicators, resulting in average **accuracy** and precision values of **82.7**% and 87.1%.

Confusion matrix - k-means *test* 2 a.)

		Empirical label	
		0	1
Cluster	0	275	59
	1	79	389

accuracy:	82,79%
precision:	86,83%
recall:	**83,12%**
specificity:	82,34%
f1 score:	84,93%

Confusion matrix - DBSCAN *test* 3 a.)

		Empirical label	
		0	1
Cluster	0	269	51
	1	79	350

0 (LLC), 1 (SLC)

accuracy:	82,64%
precision:	**87,28%**
recall:	81,59%
specificity:	**84,06%**
f1 score:	84,34%

Fig. 20 – Performance summary of tests 2 and 3, cluster vs linear model.

The smooth, intermittent, erratic, and lumpy analysis resulted in an SLC class majorly composed of erratic and lumpy items. Whereas half of the LLC class were classified as smooth and the other erratic items.

Table 8 – SIER classification per class

SIER classification

	SLC	LLC
Smooth	8,2%	47,5%
Intermittent	6,3%	0,9%
Erratic	49,9%	46,9%
Lumpy	35,7%	4,7%

Finally, the optimal silhouette scores of the DBSCAN and k-means algorithms for the studied sample were 0.47 and 0.4, respectively. This indicates that the k-means performed better than DBSCAN at generating clusters based on proximity, even though visually DBSCAN fits tighter to the dense region in the 2-dimensional space whereas k-means partitions the space homogeneously.

6 Conclusions and future remarks

Considering the proposed challenge to identify SLC items, the framework presented over the report achieves new grounds for modeling and identifying these products.

The current process for SLC identification is unstandardized and may consider business knowledge or data driven analysis of indicators. This process is of high importance for effective demand forecasting. Therefore, a standardized framework would improve this process.

The developed project considers the application of *k-means* and *DBSCAN* clustering techniques adapted to time series with precomputed *Dynamic Time Warping* distances followed by the classification of the clusters as SLC or LLC. To analyze the adequacy of this approach in the objective context, the sales were forecasted for all products and evaluated with *SMAPE*, *ME*, *MAE*, and *MSE*. Furthermore, the clusters were analyzed based on additional indicators, such as the *coefficient of variation*, *number of weeks with sales*, *intermittency*, *sales volume*, and *sales period*, were obtained for the classified clusters.

This closing chapter starts by identifying the main takeaways of the analysis undertaken and describes the relevance of the proposed approach to classify products as SLC or LLC. Finally, the future recommendations are presented in Section 6.2.

6.1 Identification of SLC items

As previously stated, SLC items are related to a wide range of common features. These products have life cycles of 13 to 26 weeks (about 6 months). However, they can reach up to 3 years. They are also characterized by spiky and intermittent demand. Traditional forecasting methods perform poorly for SLC and are usually biased. The life cycles of this group of items are marked by expansive growth followed by rapid decline.

The classes obtained with the proposed model validate these references and suggest that the model is highly adequate to divide SLC and LLC products. Based on the results presented in Chapter 5, one can conclude that the demand of the identified SLC class is highly linked to high coefficients of variation, intermittent behavior, and small number of sales. The forecast procedure for these items is characterized by high errors and biased predictions.

The overall accordance of the produced classes with the simplistic linear combination model suggests high correlation between the results of the proposed methodology and the features widely associated with SLC items in the sector.

The time series of the SLC class are typically expressed by short periods of demand defined by abrupt growth and decline, with fast trend shifts. LLC class items, however, sell through extended periods of time with stable patterns.

The clustering techniques that were tested adequately fit the sampled data. Visually, the methods intuitively separate the data points, on a partition or density basis. The optimal silhouette scores for the clustering techniques were found for models that divide the data space into two clusters, which appear to correctly correspond to SLC and LLC items.

The framework can be directly used to separate SLC and LLC items within the assortment. However, it can also be employed to train a weighted linear combination with thresholds of indicators (coefficient of variation, intermittency, number of weeks with sales, SMAPE and ME). This model can classify based on insights previously extracted from the clusters and do not require training data, leading to its efficient implementation prior to forecasting.

The implemented solution aims to reduce efforts and time spent identifying SLC items prior to demand forecasting. The adoption of this methodology in Retail Consult is motivated by the necessity of a universal framework to classify SLC and LLC items. The proven ability to accurately automate this task makes the approach appealing for this purpose.

6.2 Future remarks

Despite the improvements obtained by the conducted approach, future work can be done to further improve and refine the presented methodology.

A first step would be to improve the sample used for testing. The presented analysis was conducted to a sample of 1000 products from data relative to sales of fashion, consumer goods, books, and DIY (do it yourself) items. The presented framework can produce labels for 1000 products fast (under 50 min), however, larger samples lead to aggravated running times due to the computation cost of the DTW distance module. Improving the richness of the trained set can lead to the consolidation of other regions in the data space or further validate the ones detected. The forecasting component design is targeted to evaluate the classification model. The further development of this automated procedure can lead to refined forecasting measures and therefore better understanding of the adequacy of each model to the SLC and LLC products.

The present work bases the analysis on historical sales data, the integration of exogenous variables such as item price, season, industry status, etc., may lead to further knowledge on the identification of SLC items. As a final remark, the correct classification of new products (items with sales only registered in the last few days) is not quantified in the analysis, even though several were included in the SLC group. Understanding how the approach affects novel items classification is a ground for future research.

References

Afifi, A. A. (2020). *Demand forecasting of short life cycle products using data mining techniques*. In Maglogiannis, I., Iliadis, L., Pimenidis, E., & Jayne, C. (Eds.), Artificial Intelligence Applications and Innovations (pp. 151-162). Springer International Publishing

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). *Time-series clustering–a decade review*. Information systems, 53, 16-38.

Aitken, J., Childerhouse, P., & Towill, D. (2003). *The impact of product life cycle on supply chain strategy*. International Journal of Production Economics, 85(2), 127-140.

Bhat, H. S., & Kumar, N. (2010). *On the derivation of the Bayesian information criterion. School of Natural Sciences, University of California*, *99*.

Bulinski, J., Waszkiewicz, C., & Buraczewski, P. (2013). *Utilization of ABC/XYZ analysis in stock planning in the enterprise*. Annals of Warsaw University of Life Sciences-SGGW. Agriculture, (61 Agric. Forest Eng.).

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, pp. 226–231.

Garg, R., & Barpanda, S. (2022). *Machine Learning Algorithms for Time Series Analysis and Forecasting*. arXiv preprint arXiv:2211.14387.

HARTIGAN, John A., et al. *A k-means clustering algorithm*. *Applied statistics*, 1979, 28.1: 100-108.

He, Y., Li, J., Zhu, M., & Xu, D. (2018). *Life cycle identification and analysis of microblog hot topics*. In 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 2, pp. 156-159). IEEE.

Ho, S. L., & Xie, M. (1998). *The use of ARIMA models for reliability forecasting and analysis*. Computers & industrial engineering, 35(1-2), 213-216.

Hu, S. (2007). Akaike information criterion. Center for Research in Scientific Computation, 93, 42.

Jain, A. K. (2010). *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8), 651-666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). *Data clustering: a review*. ACM computing surveys (CSUR), 31(3), 264-323.

Kalekar, P. S. (2004). *Time series forecasting using holt-winters exponential smoothing*. *Kanwal Rekhi school of information Technology*, *4329008*(13), 1-13.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). *DBSCAN: Past, present and future*. In The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014) (pp. 232-238). IEEE.

Kheybari, S., Naji, S. A., Rezaie, F. M., & Salehpour, R. (2019). *ABC classification according to Pareto's principle: a hybrid methodology*. OPSEARCH, 56(2), 539-562.

Komninos, I. (2002). *Product life cycle management*. Thessaloniki: Aristotle University of Thessaloniki, 1-26.

Kumar, M., Patel, N. R., & Woo, J. (2002, July). *Clustering seasonality patterns in the presence of errors*. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 557-563).

Kurawarwala, A. A., & Matsuo, H. (1998). *Product growth models for medium-term forecasting of short life cycle products*. Technological Forecasting and Social Change, 57(3), 169-196.

Levitt, T. (1965). *EXPLOIT the Product Life Cycle*. Harvard Business Review, 43(6), 81-94

Li, F. (2022). *Optimization Design of Short Life Cycle Product Logistics Supply Chain Scheme Based on Support Vector Machine*. Computational Intelligence and Neuroscience, 2022.

Lopez-Molina, C., De Baets, B., Bustince, H., Sanz, J., & Barrenechea, E. (2013). *Multiscale edge detection based on Gaussian smoothing and edge tracking*. Knowledge-Based Systems, 44, 101-111.

López Oriona, Á., Montero Manso, P., & Vilar Fernández, J. A. (2022). *Time Series Clustering based on Prediction Accuracy of Global Forecasting Models*.

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

Nenni, M. E., Giustiniano, L., & Pirolo, L. (2013). *Demand forecasting in the fashion industry: a review. International Journal of Engineering Business Management*, 5(Godište 2013), 5-36.

Novia, C., Santoso, I., Soemarno, S., & Astuti, R. (2020). *Classification of product life cycle cluster to improve the performance of SMEs apple chips*. Food Research, 4(5), 1794-1801.

Ostertagova, E., & Ostertag, O. (2012). *Forecasting using simple exponential smoothing method*. *Acta Electrotechnica et Informatica*, *12*(3), 62.

Pandya, B., & Thakkar, H. (2016). *A review on inventory management control techniques: ABC-XYZ analysis. REST Journal on Emerging trends in Modelling and Manufacturing*, 2(3), 2016.

Patro, S. G. O. P. A. L., & Sahu, K. K. (2015). *Normalization: A preprocessing stage*. arXiv preprint arXiv:1503.06462.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., & Ziel, F. (2022*). Forecasting: theory and practice*. International Journal of Forecasting, 38(3), 705-871

Petrovic, S. (2006, October). *A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters*. In Proceedings of the 11th Nordic workshop of secure IT systems (Vol. 2006, pp. 53-64). Citeseer.

Raudys, A., Lenčiauskas, V., & Malčius, E. (2013). *Moving averages for financial data smoothing*. In Information and Software Technologies: 19th International Conference, ICIST

Relex Solutions. (2019). *Mastering demand forecasting of short lifecycle products*.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2013*). Gaussian processes for time-series modelling*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110550.

Şen, A. (2008). *The US fashion industry: A supply chain review*. International Journal of Production Economics, 114(2), 571-593.

Shahapure, K. R., & Nicholas, C. (2020, October). *Cluster quality analysis using silhouette score*. In 2020 IEEE 7th international conference on data science and advanced analytics (DSAA) (pp. 747-748). IEEE

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

T. W. Liao, *Clustering of time series data: A survey*, Pattern Recognit. 38 (11) (2005) 1857–1874.

TIWARI, Nidhi; VERMA, Toran. *Time Series Based Clustering Using K-Means and Hierarchical Techniques for Efficient Mining of Sales Data*.

Triana, M. J. B. (2012). *Demand forecast for short life cycle products*. Pontificia University Javeriana, Master Project. Cali/Colombia. DOI: http://vitela. javerianacali. edu. co/handle/11522/3472.