

Combined Forecasts of Intermittent Demand for Stock-keeping Units (SKUs)

AYSUN KAPUCUGİL İKİZ^a GIZEM HALİL UTMA^b

Received: 20.11.2022; Revised: 30.12.2022; Accepted: 03.01.2023


Effective inventory management requires accurate forecasts for stock-keeping units (SKUs), especially for the strategic ones for companies' operations and after-sales services like providing spare parts. Forecasting is a challenging task for such SKUs as they usually have intermittent demand (ID) patterns, consisting of many periods with zero demand and infrequent demand arrivals. Given the highly uncertain nature of ID for SKUs, this study developed a methodological framework for combining statistical and judgmental forecasts and assessed the performance of the proposed framework by using accuracy and bias measures. The forecasting process has several steps, including data preparation, data categorization based on demand patterns, generating statistical and judgmental forecasts, combining statistical and judgmental forecasts, and evaluating the forecast performance. These steps were illustrated on a real-world dataset that contains monthly customer demand data for after-sales spare parts. Results showed that combination is the best method for the majority of SKUs. This paper contributes to the limited literature by addressing the gap between the combined and ID forecasts. The proposed framework gives practitioners and researchers a comprehensive overview to help them make more accurate forecasts while encouraging the use of simple but structured approaches.


JEL codes: C44, C53, M11

Keywords: Statistical forecasting, Judgmental forecasting, Combining forecasts, Intermittent demand, Stock-keeping Units

1 Introduction

Intermittent demand (ID) is characterized by infrequent demand arrivals separated by zero-demand time intervals, and its size varies greatly depending on demand, going from thousands of units per month to a few per year. At any point in the supply chain, such demand patterns can be used to describe spare parts and stock-keeping units (SKUs), such as finished goods or semi-finished products, in the product line (Syntetos, 2001). The ID pattern is common in sectors such as aerospace, automotive, maritime, security, information technology, industrial production, and retail (Syntetos & Boylan, 2001; Ghobbar & Friend,

^a Corresponding author. Dokuz Eylül University, Faculty of Business, Department of Business Administration, İzmir, Turkey. email: aysun.kapucugil@deu.edu.tr  0000-0002-8337-2111

^b İzmir University of Economics, Faculty of Business, Department of Business Administration, İzmir, Turkey. email: gizem.halil@ieu.edu.tr  0000-0001-5040-1329

2003; Willemain et al., 2004). Also, for products that are at the end of their life cycle, this pattern typically occurs (Nagaria, 2017). Such SKUs may be fast or slow movers. Because of the intermittency of demand, companies hold a large and unnecessary amount of inventory. They might account for as high as 60% of total stock value (Johnston et al., 2003), and stock-outs (for example, spare parts in aerospace) frequently imply enormous costs, i.e., very costly operational breakdowns (Babai et al., 2019; Ghobbar & Friend, 2003). On the other, holding costs can be quite high, especially given the products' high risk of obsolescence (Saccani et al., 2017). As a result, minor improvements in forecasting demand for such SKUs may result in significant cost savings. However, due to the complex structure of these SKUs, forecasting with traditional methods is a challenging task, and they require specialized methods to generate more accurate predictions.

Several forecasting methods have been proposed to overcome the problems caused by the ID and generate more accurate forecasts. Starting with the pioneering study of Croston (1972), the literature has grown with many modifications of Croston's method (CR), such as Syntetos-Boylan Approximation (SBA) (Syntetos & Boylan, 2001), Teunter-Syntetos-Babai (TSB) method (Teunter & Duncan, 2009), and Levén & Segerstedt (2004) modification. Some nonparametric alternative approaches like Bootstrapping (Willemain et al., 2004; Hua et al., 2007; Hasni et al., 2019) and Artificial Neural Network models (Gutierrez et al., 2008; Pour et al., 2008; Kourentzes, 2013) are also proposed.

Statistical forecasting methods, on the other hand, are incapable of capturing contextual factors, such as spare parts, maintenance schedules, equipment age, and operating conditions. Even if they are based on historical data containing contextual information, they may take a while to adapt to changes in demand brought on by context-specific dynamics.

The use of contextual knowledge is essential in judgmental forecasting methods. There are two ways of using judgments in forecasting. One involves making direct forecasts based on judgment, and the other involves building forecasts using individual judgments. These two applications of judgment are used in various fields, including macroeconomic forecasting, business forecasting, political forecasting, and sports events forecasting (Parackal et al., 2007). Researchers have been paying more attention to these forecasting techniques over the past two decades (Pinçe et al., 2021). Sanders & Ritzman (1992) indicates that as the variability of the time series data increases, the judgmental forecasts generated by the practitioners may be beneficial. The accuracy of judgmental forecasts improves when the analyst has pertinent contextual knowledge, knowledge gained through experience with the forecasting environment, and up-to-date information (Lawrence et al., 2006). Otherwise, when they have limited access to quantifiable information, forecasters excessively emphasise their subjective contributions (Sanders & Manrodt, 2003; Franses & Legerstee, 2010) or repeatedly make bold adjustments based on false information (Petropoulos et al., 2016). Therefore, the judgmental forecasts may be biased and could reduce the forecast's accuracy.

Judgmental forecasts are used in three different settings. When statistical forecasts cannot be used because there is no data available, the only method left is judgmental forecasting. When data is available, a forecaster may use it to produce statistical forecasts; however, these predictions may later be subject to judgmental adjustment based on contextual knowledge. The final setting combines independent forecasts that were produced using statistical and judgmental methods. The latter is the subject of this research.

As seen from review studies (Clemen, 1989; Timmermann, 2006; Wang et al., 2022), combinations of forecasts have made significant academic strides recently, emerging as a cornerstone of forecasting research. In many cases, combining forecasts is a better approach than identifying a single best forecast, as constituent forecasts often use information from different sources. Furthermore, individual forecasts are subject to model bias from unknown model misspecifications and varyingly affected by structural breaks in the data generating process (Qian et al., 2019). These are commonly referred to as “combination forecasts” or “ensemble forecasts” (Wang et al., 2022).

The literature suggests a wide range of sophisticated combination techniques. As Li et al. (2022) stated, the simple average continually outperforms more complex weighting schemes in empirical studies like Chan & Pauwels (2018) and is still an unbeatable forecast combination technique. In the literature, this phenomenon is called a “forecast combination puzzle” (Claeskens et al., 2016; Smith & Wallis, 2009). In general, each combination method has advantages and disadvantages and which combination method should be used depends on several factors. Besides, there is still disagreement over the forecast combination technique that works best in a particular situation (Li et al., 2022; Wang et al., 2022).

In the context of forecasting ID, even though there is growing research (Pinçe et al., 2021), forecast combinations have largely gone unnoticed (Li et al., 2022). Specifically, there is very limited discussion on the combination of statistical and judgmental forecasts for ID in the literature. Therefore, given the highly uncertain nature of ID for SKUs, this research aims to develop a methodological framework for combining statistical and judgmental forecasts and assess the proposed framework’s performance by using accuracy measures. It is expected to present the overall process that connects combined and ID forecasting.

The rest of the paper is organized as follows. Section 2 presents relevant literature on combining forecasts and ID characteristics. The proposed framework for forecasting ID is described in a six-staged model in Section 3, and Section 4 explains how to combine statistical and judgmental forecasts. Section 5 illustrates the framework on real-world data that contains monthly customer demands for after-sales spare parts from an anonymous company operating in the electronics industry, manufacturing small home appliances. Finally, Section 6 concludes with the limitations of this study and recommendations for future research.

2 Relevant Literature

The section begins with a review of the literature on combining forecasts, which serves as the foundation for this study. Following that, ID patterns and associated categorization schemes are presented.

2.1 Combining Forecasts

The concept of combining several individual forecasts dates back to the 1960s when British scientist Francis Galton visited an ox weight-judging competition in which eight hundred people competed, most of whom were non-experts with varying abilities. Galton examined the 787 valid estimates provided by the contestants and discovered that the overall mean of the estimates, which represents the collective wisdom of the crowd, is nearly perfect (Surowiecki, 2005). About 60 years after Bates & Granger (1969)’s well-known work popularized the concept, a voluminous literature on forecast combinations emerged (Wang et al., 2022). After this, Clemen (1989), which summarizes a significant amount of research

from different fields, can be seen as a milestone on this topic.

Forecast combination, as a term, describes “the combination of forecasts to generate a better forecast; the component forecasts could be outcomes from model averaging, individual models, or expert forecasts” (Wang et al., 2022). Combining forecasts can be done by using some mechanical rules like averaging the included forecasts to the combination process (Lawrence et al., 2006). Clemen (1989) suggested that simply averaging the forecasts by using equal weight (i.e., $1/M$ where M denotes the number of forecasting methods to be combined) should be used as a basis when proposing more complex weighting schemes. Combination schemes range from simple methods that avoid weight estimation to sophisticated methods that tailor weights for various individual models, including model-free or model-fitting, linear or nonlinear, static or time-varying, series-specific or cross-learning, and frequentist or Bayesian (Wang et al., 2022).

Complex averaging methods have been proposed for combining forecasts, but they have yet to be successful (Green & Armstrong, 2015). Empirical evidence still shows that the simple average is much more successful (Li et al., 2022). Although the simple average can reduce forecast variance and remove uncertainty in weight estimation (Palm & Zellner, 1992), it is a sensitive statistic in extreme values. As a result, other strong combinations utilizing the median and trimmed means have received some attention (Petropoulos & Svetunkov, 2020). The organizers of the M5 competition, centered on sales forecasts using a large number of intermittent time series, recently used the simple average of exponential smoothing and ARIMA as the combination benchmark, which performed better or equally well as the individual methods that constituted them (Makridakis et al., 2022).

As each combination method has its own merits, which combination method should be used depends on the kind of forecasts (deterministic, probabilistic, quantile, etc.), the size and quality of the model pool, the available information, and the particular forecasting issues (Wang et al., 2022). However, there is still disagreement over the forecast combination technique that works best in a particular situation (Li et al., 2022; Wang et al., 2022).

When the constituents of forecast combination come from judgmental and statistical methods, techniques that combine these two categories of forecasts include their simple average, judgmental bootstrapping (i.e., a type of expert system that converts expert reasoning into a set of explicit rules, Armstrong, 2001a), and statistical techniques that aim to eliminate systematic biases from judgemental forecasts (Parackal et al., 2007).

Among studies using these techniques, Lawrence et al. (1986) investigated the effectiveness of combining forecasts for time series with various forecasting difficulty and seasonality levels. They found out that for series with low MAPE, the combination is more effective and seasonality has no influence on the benefits of the combination. Working on the effects of difficulty levels of series with a coefficient of variance instead of MAPE, Sanders (1992) also found out that combining forecasts is most effective for simpler series. For harder series, combined forecasts are less accurate than judgmental forecasts because judgments can generate better forecasts than statistical models. Weinberg (1986) worked on the forecasts for attendance at events. In his study, by using an econometric model, *ex-ante* forecasts were generated. Also, managers generated forecasts by using their judgment. Results showed that the econometric model provided more accurate results than managers’ judgment. However, the accuracy of the combination of the econometric model and judgment was superior to both model and judgment alone.

There are some studies showing that adjusting the statistical forecast by using judgments increases the accuracy in the forecasting of ID (e.g., Syntetos et al., 2009; Davydenko & Fildes, 2013). However, adjustments are subject to biases and may harm the forecast accuracy (Eroğlu & Croxton, 2010). Mechanically integrating statistical and judgmental methods should be preferred to avoid such biases (Sanders & Ritzman, 2001).

The combination of statistical and judgmental forecasts for ID has rarely been discussed in the literature, such that only the simple averaging method has been used to improve the forecasting (Petropoulos & Kourentzes, 2015). Especially for highly ID patterns, Bates & Granger (1969) found that using weighted forecast combinations causes the covariance matrix of forecast errors to be singular. As the obtained errors may contain a large number of zero values, it is impossible to calculate this matrix's inverse. A similar issue is also valid for regressive approaches. Thus, many sophisticated forecasting methods are inapplicable due to the numerous zeros and non-smooth patterns present in ID.

2.2 Intermittent demand patterns

Syntetos (2001, p. 365) stated that “infrequent demand occurrences and variable demand size when demand occurs mean demand to be non-normal since demand per unit period or lead time demand cannot be represented by the normal distribution”. The literature proposes various forecasting methods for different non-normal demand patterns. Specifying and categorizing the demand patterns according to their similarities is essential to find the best-performing forecasting method, which provides the highest forecasting accuracy.

Williams (1984) proposed to categorize the patterns based on the idea of variance partition of the demand during lead time as “variance of the demand sizes”, “transaction variability”, and “variance of the lead times”, and classified the SKU demand into three categories: smooth, slow-moving, and sporadic. After this study, several authors proposed different categorizations to determine the best-performing forecasting methods and inventory control parameters.

Two parameters -“the average inter-demand interval” (p) and “the square of the coefficient of variation” (CV^2)- are used to determine the characteristics of demand data. p shows the regularity of demand by measuring the average number of periods between two non-zero demands, and it is calculated as follows.

$$p = \frac{\sum \text{Intervals between non - zero demand periods}}{\text{Number of non - zero demand periods}} \quad (1)$$

The coefficient of variation (CV), on the other hand, measures the variation in the demand size and is calculated as follows.

$$CV = \frac{\text{Standard deviation of demand values}}{\text{Average demand over periods}} \quad (2)$$

Johnston & Boylan (1996) showed that Croston (CR) method outperforms the Exponentially Weighted Moving Average (EWMA) for p values greater than 1.25 review periods and mentioned that these SKUs have ID patterns. The authors were the first ones to formally confirm the importance of the value of p as a classification parameter (van Kampen et al., 2012). Eaves (2002) reclassified the SKUs in the dataset by modifying Williams (1984)'s classification as smooth, irregular, slow-moving, mildly intermittent, and highly intermittent. However, as the cut-off values of this categorization are solely dependent on

the properties of the underlying dataset and adequate subsample size considerations, it is not widely applicable.

Syntetos et al. (2005) compared EWMA, CR, and Syntetos-Boylan Approximation (SBA) methods based on the theoretical analysis of the Mean Square Error (MSE) to determine the regions of superior performance and define the demand patterns accordingly. They developed a model of four demand categories: “erratic”, “lumpy”, “smooth”, and “intermittent”, as shown in Figure 1. The categorization scheme proposed by the authors is based on p and the CV^2 of demand sizes. Comparing the theoretical MSE values of EWMA, CR and SBA, the cut-off values are determined as $p=1.32$ and $CV^2=0.49$. Syntetos-Boylan-Croston (SBC) scheme was empirically tested using 3,000 SKUs from a company operating in the automotive industry, and the validity is confirmed. Syntetos et al. (2005) contributed to the identification of CV^2 as a new categorization parameter for demand forecasting purposes.

Lastly, Kostenko & Hyndman (2006) developed the KH scheme as an extension of the SBC scheme, which is more accurate and simpler. The authors suggested using the SBA method in a smooth pattern if $(CV)^2 > 2 - (3/2)p$. According to the KH categorization scheme, for SKUs with CV^2 value of 0.4 and p value of 1.25, SBA is used, while the SBC categorization scheme suggests using CR for the same SKUs.

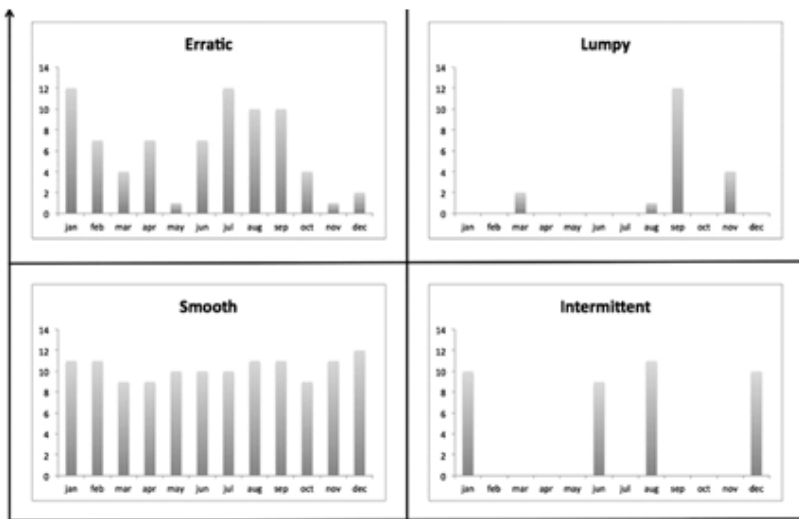


Figure 1: Demand Patterns
Source: Constantino et al. (2018, p.59)

SKUs with smooth demand patterns show a regular demand over time, and the non-zero demand size shows little variation. Infrequent demand occurrences may be defined as intermittent (or sporadic) demand. Intermittence refers to the occurrence of the demands but not the sizes of the occurring demands. For this type of demand, the average time between consecutive transactions is significantly longer than the unit period, the latter being the period for updating forecasts (Silver et al., 1998). In the case of an erratic demand pattern, demand for SKUs is highly variable and unstable. In lumpy demand, there are many periods without any demand occurring, and when the demand occurs, its size varies. Since both demand size and demand occurrences are highly variable, it is challenging to forecast SKUs with lumpy demand (Syntetos, 2001).

3 Methodology

The purpose of this study is to present a methodological framework for combining statistical and judgmental forecasts in order to obtain more accurate results for SKUs with ID patterns. The proposed framework for forecasting ID is described as a six-stage model, as shown in Figure 2.

In the first stage, the data is prepared for the forecasting process. The second stage is categorizing the demand data based on demand patterns (intermittent, smooth, lumpy and erratic) mentioned in Section 2.2 by using the categorization schemes proposed for the SKUs with ID patterns. In the third stage, a statistical forecasting model is built using the appropriate ID methods. The best-performing forecasting method is selected using accuracy measures appropriate for ID data. Parallel to statistical forecast model building, the judgmental forecasting model is also built in the fourth stage. The fifth stage is where the best statistical forecasts are combined with the judgmental forecasts by using combining procedures. Finally, the accuracies for statistical, judgmental, and combined forecasts are calculated, and the best-performing methods are selected for each SKU in the sixth stage. The following sections explain each stage of this framework.

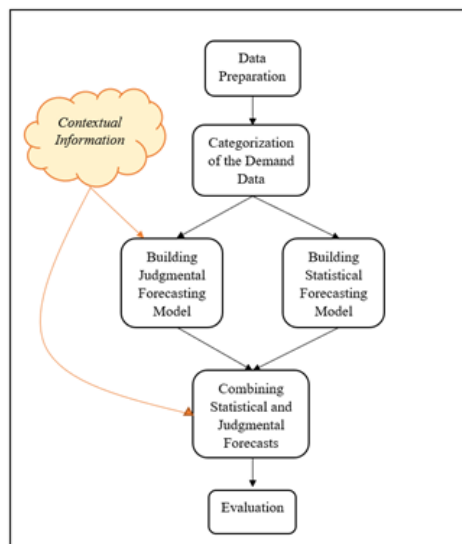


Figure 2: Stages of the proposed methodological framework

3.1 Data preparation

The first stage is to prepare the data for building a forecast model. The data usually gathered from companies may not be suitable for directly using in the forecasting process. For instance, historical data for an SKU may have missing values, consisting only of zeros or a high number of consecutive zeros, especially in the beginning or ending periods, which may indicate that SKU is a new product or it is not in use anymore. Such SKUs should be eliminated from the dataset to start the forecasting process.

3.2 Categorization of the demand data

The second stage is the categorization of the demand data. Categorizing the data according to the SKUs' demand patterns plays an essential role in selecting the best-performing forecasting method and inventory control parameters. Companies hold high numbers of SKUs, so it is suggested that, instead of evaluating them on an individual basis, it is more effective to assess them as groups with similar characteristics. In the literature, there are several categorization schemes proposed by various researchers (e.g., Williams, 1984; Johnston & Boylan, 1996; Syntetos, 2001; Ghobbar & Friend, 2002; Eaves, 2002; Syntetos et al., 2005; Kostenko & Hyndman, 2006; Boylan et al., 2008).

In this study, the Syntetos-Boylan-Croston (SBC) scheme proposed by Syntetos et al. (2005) is used since the majority of the studies reported accurate results based on this scheme (Fildes et al., 2019), and it is also easy to understand and interpret from the point of the users. The SBC scheme uses average inter-demand interval (p) and squared coefficient of variation (CV^2) to categorize the demand patterns into four groups: smooth, intermittent, erratic, and lumpy. The authors defined the demand patterns and established the regions of superior performance by comparing the theoretical MSE values of EWMA, CR and Syntetos-Boylan Approximation (SBA). The SBC scheme was empirically tested by using 3,000 SKUs from a company operating in the automotive industry, and the validity is confirmed. The results showed that the cut-off points are $p=1,32$ and $CV^2=0,49$. Figure 3 visualizes the categorization scheme recommended by Syntetos et al. (2005).

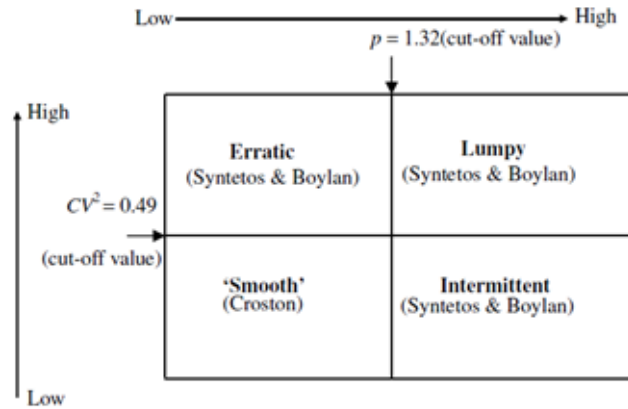


Figure 3: SBC categorization scheme
Source: Boylan et al. (2008, p.476)

Based on the cut-off values of SBC categorization, p , and CV^2 of each demand pattern, SBS categorization suggests that forecasting the smooth demand with CR gives the highest accuracy while erratic, lumpy, and ID are forecasted best by using SBA.

3.3 Building a statistical forecast model

After examining the data patterns using an appropriate categorization scheme, in the third stage, statistical forecasts should be generated using the candidate models for each category and evaluated based on their accuracies to select the best-performing one.

ID patterns can be modeled by using several statistical forecasting methods (Waller, 2015). This study covers three widely used methods in the literature: the single exponential smoothing (SES) as a traditional method, the pioneer Croston method (CR) proposed for ID pattern and Syntetos-Boylan Approximation (SBA) as one of the CR's variants.

3.3.1 Single Exponential Smoothing

In practice, Simple Exponential Smoothing (SES) is commonly used due to its straightforwardness and robustness. It is a parameter-based method. It provides an EWMA of all observed values. This method aims to estimate the current level to use when forecasting future values. SES is appropriate for data without a seasonal pattern or trend. SES revises an estimate by using more recent experiences. The calculations require a predetermined parameter, “ α ”, the so-called smoothing constant. The most recent observation receives the highest weight, and the older observations receive less weight. The smoothing constant is a value between 0 and 1 and is determined judgmentally and based on the data's characteristics. Silver et al. (1998) suggest choosing a value for the smoothing constant between 0.1 and 0.3 if forecasting is done monthly. The general formulation of exponential smoothing is

$$Y_{t+1} = \alpha A_t + (1 - \alpha) Y_t \quad (3)$$

where Y_t and Y_{t+1} are the old and new smoothed value of the forecast for periods t and $t + 1$, respectively, and A_t is a new observation or the actual value of the series in period t .

SES is a widely used method by practitioners for SKUs that have both smooth and non-smooth demand patterns. As a strong candidate among conventional time-series forecasting techniques, SES is chosen for testing its effectiveness in the context of ID forecasting.

3.3.2 Croston's Method (CR)

Croston (1972) proved that using SES is inadequate for forecasting the ID due to the biases it causes. Syntetos et al. (2015, p. 1747) stated that “In SES, data that is more recent weights more heavily. Thus, just after a demand occurs, it gives forecasts that are biased high while it gives forecasts that are biased low just before a demand”. This situation results in high replenishments and excessive stock levels.

In order to address this problem, a new forecasting method, the CR model, which separately estimates the non-zero demand size and inter-arrival time between subsequent demands by using SES, is proposed (Croston, 1972). CR is the first proposed method, especially for SKUs that have ID patterns. It assumes that demand occurs as a Bernoulli process; intervals between demands are independent and identically distributed, and demand sizes are independent and normally distributed. According to the CR, the ratio of estimates of the mean size of non-zero demand (Z_t) to the mean interval between non-zero demands (P_t) provides an estimate of the mean demand per period, Y_t , as follows.

$$Y_t = \frac{Z_t}{P_t} \quad (4)$$

The algorithm for CR, which is provided in Table 1, estimates also uses the observed value of the demand, A_t , and the time interval since the last demand, Q .

Table 1: Croston Method's Algorithm

If $A_t = 0$	If $A_t \neq 0$
$Z_t = Z_{t-1}$	$Z_t = \alpha A_t + (1 - \alpha) Z_{t-1}$
$P_t = P_{t-1}$	$P_t = \alpha Q_t + (1 - \alpha) P_{t-1}$
$Q = Q + 1$	$Q = 1$

Figure 4 schematizes the CR forecasting process. When there is not any demand in a review period, the estimates are not changed. If the demand is zero in period t , the algorithm only increments the count of periods since the last positive demand (Croston, 1972). The forecasts are updated only after the occurrence of positive demand. If demand occurs every period, CR's forecasts are the same as forecasts that SES generate. So, it is possible to use this method both for intermittent and smooth demand patterns.

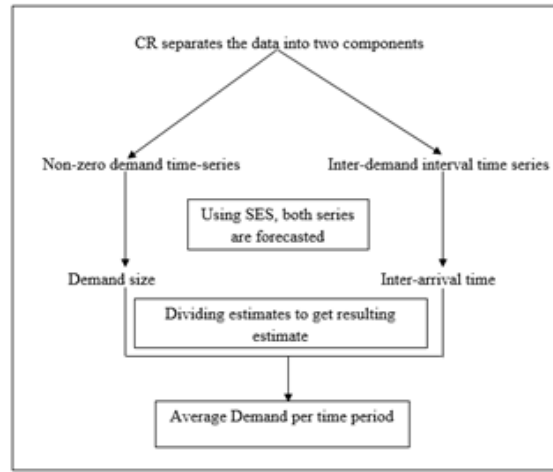


Figure 4: Forecasting process of Croston's method
Source: Nagaria (2017)

Several studies showed the superiority of CR over traditional methods (Willemain et al., 1994; Johnston & Boylan, 1996; Ghobbar & Friend, 2003). Due to its proven success in forecasting the demand for the SKUs with ID and its applicability without any additional costs for the organizations, this method is also used while generating statistical forecasts.

3.3.3 Syntetos and Boylan Approximation (SBA)

Due to some mathematical derivation problems, Syntetos & Boylan (2001) reported that CR is biased. Thus, Syntetos-Boylan Approximation (SBA) was proposed as a revised model, which approximately corrects the bias in CR's demand estimates. The new mean demand estimator is given as follows.

$$Y_t = \left(1 - \frac{\alpha}{2}\right) \frac{Z_t}{P_t} \quad (5)$$

SBA is the most widely used variant of CR. Several studies in the literature showed that SBA outperforms other methods such as SES, CR, and MA (Eaves & Kingsman, 2004;

Syntetos & Boylan, 2001; Syntetos et al., 2005; Gutierrez et al., 2008). This method is selected since it improves the quality of forecasts generated by CR.

Another issue is that neither of these methods causes any additional cost for the companies since it is easy to apply by using simple software like MS Excel or RStudio. Due to the success and prevalence of these methods both in literature and practice for ID forecasting, they are selected for this study.

3.3.4 Performance measures of Statistical Forecast Model

In practice, the generated forecasts from a forecast model are rarely perfect. Forecast accuracy is the most important criterion when deciding whether to use a forecasting method or not. The next step in the statistical forecasting process is the measurement of forecast accuracies by using appropriate measures to see the performance of each method and related parameters. These measures are based on the forecast error, e_t , the difference between actual value, A_t , and forecast, F_t , for a given period t as follows.

$$e_t = A_t - F_t \quad (6)$$

The e_t values closer to 0 show that the forecast is close to the actual. Positive e_t value shows that the forecasts are smaller than the actual (underestimation), whereas negative e_t values indicate an overestimation in the forecasts, i.e., forecasts are larger than the actuals.

The error measures are categorised into scale-dependent, scale-independent, percentage, and relative. Scale-dependent errors have the same scale as the data and are used for comparing different forecasting methods applied to the same data. It is not appropriate to use such measures to compare datasets with a different scale. In scale-independent error measures, the error is scaled and becomes independent of the scale of the data. They can be used to compare forecasting methods both on a single series and between different series. Relative error measures are used to determine the best-performing method by dividing the forecast errors obtained by using different forecasting methods from each other. This division scales the measures (Hyndman & Koehler, 2006).¹

As the chosen forecast error measure can influence the performance ranking of forecasting methods, there is no metric that was universally best (Silver et al., 1998). The most appropriate performance measures for intermittent series are mean absolute scaled error (MASE, as the standard measure for the data with different scales and zero values, Hyndman & Koehler, 2006), scaled mean absolute error (sMAE), scaled mean square error (sMSE), scaled cumulative error (sCE, as traditional (accuracy) measurements, Wallström & Segerstedt, 2010), and scaled periods in stock (sPIS, as bias error measure, Hyndman & Koehler, 2006; Kourentzes, 2014; Petropoulos & Kourentzes, 2015). Relative Geometric Root Mean Square Error (RGRMSE) is also one of the most used relative error measurements first proposed by Fildes (1992). Syntetos (2001) and Syntetos & Boylan (2001) recommended RGRMSE as it is a well-behaved accuracy measure to use in ID. In the last M5 competition, Root Mean Squared Scaled Error (RMSSE) was required to evaluate the accuracy of point forecasts for ID (Makridakis et al., 2022). After calculating the forecast accuracies for each SKU, the results are compared, considering different measures. The final forecasts are generated using the best-performing methods and parameters for each SKU.

¹ The details of the error measures are available in Appendix.

3.4 Making judgmental forecasts

The mental capacity of humans for processing information has reasonable boundaries. Judgmental forecasters quickly arrive at a point where more information is no longer useful in making more accurate forecasts. Furthermore, regardless of how intelligent humans are, they are incapable of learning about complex relationships solely through experience. Thus, it is difficult for them to forecast complex, uncertain situations without the aid of structured techniques (Green & Armstrong, 2015). In general, it has been demonstrated that standardizing the techniques applied by experts will increase accuracy (Armstrong, 2001a).

The jury of executive opinion, the Delphi method, analogies, and scenario forecasting are some of the judgmental methods that can be used to reflect experience and knowledge about the SKUs in the process of forecasting ID. These techniques, which are described in more detail below, can be used to produce judgmental forecasts in parallel with the statistical model construction in Stage 3 of the process.

3.4.1 The jury of executive opinion

The jury of the executive opinion, as one of the most common methods used in judgmental forecasting, relies on the opinions and expertise of high-level managers, executives, or experts who have the best insights about the firm's future situation. The final group forecast becomes the forecast for a product as a blend of opinions of these executives from different functional areas such as sales, marketing, and production. The opinions can either be collected by using personal interviews or group meetings.

Though in group meetings, there is a chance of discussion of various viewpoints, there may also be the domination of an expert with a strong personality which may affect the final forecast (Wilson & Keating, 2008). In the jury of executive opinion method, companies can also use statistical models to help with the analysis (Wright, 2013).

3.4.2 Delphi method

Assuming that a group's forecast is more accurate than an individual's forecast, the Delphi method aims to construct consensus forecasts from a group of experts in a structured and iterative manner (Hyndman & Athanasopoulos, 2018). The Delphi procedure, implemented and managed by a facilitator, has the following steps: (i) Assemble a panel of experts between 5 and 20 with diverse expertise. (ii) Set the forecasting tasks and deliver them to the experts. (iii) Experts provide preliminary forecasts and rationales and summarize the initial forecasts to provide feedback to the experts. (iv) Deliver feedback to experts so that they can revise their forecasts based on the opinions of others. This process is repeated two to three times till the experts reach a satisfactory level. (v) By aggregating the forecasts of experts, construct the final forecasts.

This procedure has four key components: anonymity, iteration, controlled feedback, and group response aggregation. During the processes, all participating experts maintain their anonymity and express their opinions privately without social and political pressure. When using group-based processes to gather and synthesize the information, anonymity can minimize the effects of dominant individuals, which are typically a problem (Dalkey, 1969). The geographical dispersion of the experts, as well as the use of electronic communication, facilitates confidentiality and reduces problems associated with group dynamics, such as manipulation or duress to accept a viewpoint (Hsu & Sandford, 2007). Iteration allows

experts to change their minds without losing face in the eyes of the remaining group members (Hanke & Wichern, 2014). Usually, two or three iterations are sufficient since the experts may drop out as the number of iterations increases (Hyndman & Athanasopoulos, 2018). The Delphi method uses controlled feedback to reduce the effect of noise which refers to group interaction that both affects the data and deals with other expectations rather than concentrating on the main topic. With controlled feedback, a structured description of the prior iteration is distributed to the experts. Using the feedback, experts can generate additional insights (Hsu & Sandford, 2007). Group response is usually generated by giving each expert's forecast equal weight.

3.4.3 Forecasting by analogies

One of the judgmental forecasting methods is to use analogies. Analogy contains information about how people behaved in a similar situation in the past. Forecaster identifies a pattern that happened in the past and applies the same pattern to a new issue. It is expected that analogies will be helpful in forecasting decisions in conflict situations, like strikes or international disputes, since they provide essential information for difficult situations to forecast (Green & Armstrong, 2007). However, analogies are generally used in an unstructured way when people make judgmental forecasts. Armstrong (1985) showed that structured judgmental forecasting methods provide higher accuracy than unstructured ones.

The structured analogy is related to Case-Based Reasoning (CBR), which is used in cognitive science and artificial intelligence. In CBR, information about situations (cases) is stored with the intention of recalling cases that are comparable to a target problem assisting in problem-solving (Armstrong, 2001a). Green & Armstrong (2007) proposed a structured approach for forecasting with analogies, which may encourage experts to consider more information on analogies and to process it effectively.

Structured analogies have five steps: (i) The appointed administrator describes the target situation briefly and accurately. (ii) The administrator chooses at least five experts who have knowledge about similar situations to the target situation. (iii) Experts identify and describe as many analogies as they can. Based on each of these analogies, forecasts are generated. Green & Armstrong (2007) stated that experienced forecasters on analogies who have more than two analogies generate the most accurate forecasts. (iv) Experts list similarities and differences between their analogies and target situations. Later they rate the similarity of each analogy to the target situation on a scale. (v) From experts' analogies, the administrator drives the forecasts using a set rule. This rule can be a weighted average where the weights can be guided by the ranking scores of each analogy by the experts (Hyndman & Athanasopoulos, 2018). The lack of situations comparable to the target is a limitation of structured analogy.

3.4.4 Scenario Forecasting

Scenario-based forecasting is fundamentally different from judgmental forecasting in its methodology. This method creates forecasts using scenarios that are intended to be plausible but not necessarily most likely. These scenarios either portray a quick snapshot of the future or a believable progression from the present to the future (Bunn & Salo, 1993). Each scenario-based forecast may have a low chance of occurring, unlike the Delphi and using an analogy, where the anticipated outcome is meant to be a likely one.

The effects and interactions of all potential factors and forecasting goals are considered when creating the scenarios. The scenarios help managers to understand the role of uncertainties better. Also, some extremes, like the best, middle, and worst-case scenarios, can be identified when building forecasts based on analogies. Keeping track of these extremes can help with early emergency planning (Hyndman & Athanasopoulos, 2018). In capital-intensive industries like oil companies, vehicle manufacturers, and electric suppliers, which have long planning horizons, scenario techniques are found to be more popular.

The selection of judgmental procedures is influenced by significant changes, frequent forecasts, disagreements among decision-makers, and policy considerations (Armstrong, 2001b). If the expected changes are not significant, methods are likely to differ slightly in accuracy. Expert forecasts, which can be adapted to the condition and prepared instantly, may also suffice for infrequent forecasts. If decision-makers expect large changes in the situation and are not in conflict, forecasts can be obtained from experts through the jury of executive opinion or the Delphi method. Scenario forecasting is an alternative when decision-makers need forecasts to examine different policies, and it is difficult to find relevant analogies.

4 Combining statistical and judgmental forecasts

The fifth stage combines the forecasts from the best statistical model and the judgmental forecasts. Combining forecasts generated by using different methods plays an essential role in improving overall accuracy. Each technique adds different information to the forecasting process, which a single technique is not capable of. Blattberg & Hoch (1990) stated that the final forecasts generated by combining judgmental and statistical forecasts are more accurate than the constituent forecasts. The opinions of the experts add contextual information, which explains the unexpected issues in the data. Also, it helps forecasters to understand future events which would affect the historical data.

Combination procedures can range from mechanical methods, such as taking a simple or weighted average of the constituent forecast, to using judgment to determine how forecasts should be combined (Lawrence et al., 2006). The simplest approach is using an equal arithmetic average of the individual methods (i.e. $1/M$, where M is the number of forecasting methods to be combined), which provides improvements in accuracy for many forecasts (Clemen, 1989). More complex weighting schemes can also be determined by judgments based on which modeling approach seems strongest or by several trials considering equal or alternate weighting. For instance, if the practitioner believes that there are some issues that may cause one constituent forecasting method to perform better than the other, s/he can weigh this method heavier. However, this complexity and the time waste it causes would make combining forecasts less desirable for an organization (Sanders & Ritzman, 2004). Forecast errors range from 5.5% to 94% when forecasts are combined using a complex method (Duncan et al., 2001). The findings of Fildes & Petropoulos (2015) supported differential weighting in situations when there is prior evidence on which methods provide forecasts that are most accurate given the conditions.

Armstrong (2001c) proposed some procedures for combining the forecasts. Combining should be done mechanically to obtain greater accuracy and reduce bias, and all procedures should be described in more detail. When using judgment, it should be done in a structured manner, and the details of the procedure should be documented. When there is insufficient information about the relative accuracy of alternate forecasting sources, judgmental

weights should be avoided. When subjects provide positive feedback about the accuracy of the sources, judgmental weighting is more accurate (Fischer & Harvey, 1999). Armstrong (2001c) stated that when the practitioner is uncertain about the forecasting methods to be used, each forecast should be weighted equally. Fildes & Petropoulos (2015) found that differential weighting is appropriate when there is prior knowledge of the methods that produce forecasts that are the most accurate given the circumstances.

4.1 Evaluation

The last stage of the framework is evaluating the performances of statistical, judgmental, and combined forecasts to find out the best-performing one for each SKU.

In the final evaluation, the performance of each model is compared based on MASE, RMSE, S-MAPE, and bias in this study. Bias is calculated by averaging the difference between the actual and predicted value. This value should be close to zero if the forecast is unbiased. The positive value refers to underestimation, and the negative value means that the forecasting method overestimated the values.

Aside from producing the most accurate forecasts, the chosen forecasting method should also produce forecasts that are timely and easily understood by management, allowing the forecast to aid in making better decisions (Hanke & Wichern, 2014).

5 Empirical Illustration

The proposed framework of this study is illustrated on real-world data that contains monthly customer demands (60 months, from 2014 to 2018) for after-sales spare parts from an anonymous company operating in the electronics industry, manufacturing small home appliances. The company did not provide any information about details except the amounts of the demanded SKUs, and all SKU names were changed for anonymity. 8,498 SKUs are presented in a raw form in this dataset.

In this study, it is assumed that the SKUs in the dataset are independent of one another because the company withheld information about the SKUs due to anonymity. Furthermore, it is assumed that the historical data was accurately and properly recorded.

5.1 Data preparation

The demand data is examined to prepare for the forecasting process. All calculations and data analysis is performed by using MS Excel 2019 and RStudio (version 3.6.0).²

First, the historical demand dataset is examined to determine whether it is incomplete or inconsistent. Some SKUs that have not been demanded in the last 60 months are being phased out. Also eliminated are SKUs with no or too few demand occurrences in the first and last 24 months, those with less than ten demand occurrences in 60 months, and products with no demand between months 24 and 36. Finally, there are 2,431 SKUs left, totaling 145,860 data points. Descriptive statistics given in Table 2 show the variety in the demand intervals, non-zero demand sizes, and demand per period. The demand intervals imply that the number of consecutive zero demands is high.

² The codes are available upon request.

Table 2: Descriptive statistics for after-sales spare parts

2,431 SKUs in Total	Demand intervals		Demand sizes		Demand per period	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Minimum	1.00	0.00	1.11	0.79	0.63	6.24
First quartile (25%)	1.07	0.26	5.24	19.51	5.39	28.61
Median	1.52	0.85	9.84	37.24	10.65	48.10
Third quartile (75%)	2.63	2.07	17.83	64.47	19.05	75.09
Maximum	7.30	4.92	44.69	149.37	50.19	169.44

5.2 Categorization of the demand data

According to the SBC categorization scheme, considering the average inter-demand interval and squared coefficient of variation values for spare parts, the dataset is categorized by using the RStudio’s “tsintermittent” package (Kourentzes & Petropoulos, 2016) and the idclass function. 389 SKUs have smooth demand patterns, while the remaining 2,042 SKUs have erratic, lumpy, and ID patterns, as shown in Figure 5.

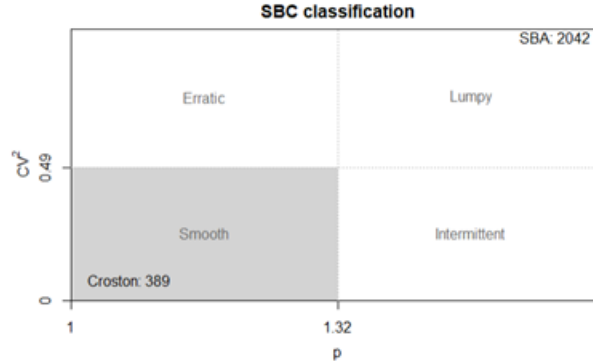


Figure 5: SBC classification for after-sales spare parts

In Table 3, the distribution of the dataset according to demand categories is presented using the p and CV^2 values for classification. The majority of the SKUs in the dataset have ID, while the least number of SKUs (16%) have a smooth demand pattern.

Table 3: Demand patterns for after-sales spare parts

Pattern	Condition	Quantity	Percentage
Smooth	$p \leq 1.32$ and $CV^2 \leq 0.49$	389	16.0
Erratic	$p \leq 1.32$ and $CV^2 > 0.49$	582	23.9
Intermittent	$p > 1.32$ and $CV^2 \leq 0.49$	802	33.0
Lumpy	$p > 1.32$ and $CV^2 > 0.49$	658	27.1

5.3 Building statistical forecast model

This stage is primarily concerned with obtaining the best statistical forecast model for monthly customer demands for after-sales spare parts. The analysis is carried out in the following steps to obtain this model: identification of statistical forecasting methods based on the nature of the data, estimation of models, evaluation of models based on forecast accuracy, and selection of the final model to generate forecasts.

First, the statistical forecasting methods that will be applied to each SKU in the dataset are determined as the traditional single exponential smoothing (SES), the pioneer Croston method (CR) proposed for ID pattern, and Syntetos-Boylan Approximation (SBA) as one of the CR's variants.

Each model is estimated in the second step using the proper parameters. The datasets used to build forecast models are typically split into the training and test sets (Hyndman & Athanasopoulos, 2018). Out of 60 months, the first 42 are used as a training set to estimate the statistical forecasting model's parameters, and the remaining months are used as the test set to evaluate the model's accuracy. The "tsintermittent" package of RStudio is used to develop the models. The package offers "sexsm" function for building the traditional SES model and "crosst" function for the CR and SBA models. These three methods use the smoothing constant (α) for forecasting as well as initial values to start their algorithms, so the selection of these parameters will directly affect how well they perform.

The literature usually suggests a small α ranging from 0.05 to 0.30. For ID data, low smoothing constants between 0.10 and 0.20 are realistic (Croston, 1972; Syntetos et al., 2005). Therefore, α values are chosen between 0.05-0.30, incrementing in 0.05 in this analysis. Regarding the initial values to start their algorithm, SES requires one initial value for the first estimate, whereas CR and SBA require two initial values (separately for demand size and the interval between non-zero demands estimates). These initial values can be determined by using the naïve method (one step ahead), or calculated by taking the mean of the related values, or determined by the practitioner judgmentally.

Similarly, practitioners may choose to use optimal parameters rather than fixed initial values. Cost (or loss) functions are used for the optimization process, and they measure the fit of a model to the actual data. MAR (mean absolute rate), MAE (mean absolute error), MSR (mean squared rate) and MSE (mean squared error) are the cost functions used in this analysis, and the parameters minimizing the specified cost functions are chosen.

The "crosst" function offers an option to use single or two optimal α parameters for CR and SBA. The parameter α can be calculated automatically by choosing one of the predetermined cost functions. In this study, besides the preset α values, forecasts are generated by using all combinations of single and double α values that optimize each cost function. For the preset forecasts, the calculations are repeated for the mean and naïve initial values. In total, for SBA and CR, 56 different forecasts are generated for each SKU. The forecast horizon is 18 months which is equal to the length of the test set.

Once the forecasts are generated, the best-performing method and parameters for each SKU are chosen based on the forecast accuracies in the third step. For evaluating the statistical forecasting methods, scaled mean absolute error (sMAE), scaled mean squared error (sMSE), root mean squared error (RMSE), symmetric mean absolute percentage error (S-MAPE), mean absolute scaled error (MASE), scaled cumulative error (sCE), and scaled periods in stock (sPIS) are used in the study. Scaling the errors turn the accuracy measure into a scale-independent form. Scale-independent errors can be compared across the series. That is why sMAE, sMSE and MASE are selected. Also, MASE is commonly used in evaluating ID forecasts since it is appropriate to use for this problematic demand pattern. RMSE is a scale-dependent error measure, and it is widely used in practice. So, it is included in the comparison process as well. S-MAPE shows accuracy as a percentage. Also, it overcomes the division by zero problems that MAPE has in the ID context.

These measures are not capable of determining whether there is a systematic error, bias, or not. To determine if the forecasting methods generate results which are not biased, it is essential to use bias measures together with other accuracy measures. Since sCE and sPIS are scaled measures, they are used as bias measures to make comparisons across the series.

The forecast accuracies are obtained by using “smooth” and “Metrics” packages of RStudio. “Accuracy” function of “smooth” package provides accuracy measures as sMAE, sMSE, sCE and sPIS while the MASE, S-MAPE and RMSE values can be calculated with the help of “Metrics” package.

Table 4: Accuracy and bias measures for the sample SKU-P1 (partial representation)

MODEL	MASE	SMAPE	RMSE	sMAE	sMSE	sCE	sPIS
P1.acc.m.ses020	2.07	0.87	808.02	0.75	0.52	7.47	54.97
P1.acc.m.ses025	2.43	0.93	914.73	0.88	0.66	8.76	63.38
P1.acc.m.ses030	2.82	0.99	1,035.80	1.02	0.85	10.19	72.63
P1.acc.m.sba005	1.05	0.72	429.07	0.38	0.15	1.49	16.09
P1.acc.n.sba010	1.12	0.69	518.04	0.40	0.21	3.44	28.79
P1.acc.n.sba015	1.19	0.70	549.37	0.43	0.24	3.96	32.15

All these accuracies are retrieved from the relevant RStudio functions and then organized in an Excel file shown partially in Table 4 for the evaluations. This accuracy table has 136.136 rows and seven columns. The first column (Model) shows an ID which refers to a description of the model performed. This ID is formed as having five components: SKU number (e.g., P1), accuracy function (e.g., acc), initial value (m for mean, n for naïve and o for optimum), forecast method (CR, SBA or SES) and α value (0.05-0.30).

In the last step, the accuracies of forecasting methods for each SKU are controlled. Based on different measures, the best-performing method and parameters are selected (i.e., α and initial values) for the demand dataset. For example, in Table 4, the best-performing forecasting model for the SKU coded as P1 is the SBA method with an α value of 0.05 and the initial values determined by the mean. This selection is made based on considering all of the accuracy measures since the best-performing method is not always the same for all of them. Except for sCE and sPIS, smaller values show that the forecast is more accurate. For sCE and sPIS, which are bias measures, values that are close to 0 are preferable. Otherwise, their positive or negative values indicate stock problems.

The results are presented overall based on the demand pattern of the SKUs. The number of SKUs that perform best for each forecasting method and parameter is used to assess the performance of these methods and parameters.

Based on all preset smoothing α values and those calculated by minimizing the cost functions, the statistical forecasting models performed best for the majority of all demand patterns when the α parameter is set at 0.30 (Table 5). Opposite to the suggested guidelines in the literature, larger smoothing values performed better for most SKUs in this dataset.

To identify the best method for determining the initial values for selected statistical forecasting models, the number of SKUs that performed best for each of the mean value, the naïve estimate, and the optimum value obtained from the optimization process is calculated. For the majority (41.2%; 1,002 out of 2,431) of SKUs that have erratic, intermittent, and lumpy demand patterns, using the mean of a few more recent observations in the dataset as the initial values of statistical forecasting models resulted in better performance in terms of accuracy (Table 6). For a smooth demand pattern, the naïve method is better for deter-

mining the initial values.

Table 5: Optimum smoothing parameters based on the demand pattern

α	Erratic (582)		Lumpy (658)		Intermittent (802)		Smooth (389)	
0.05	41	7%	88	13%	145	18%	54	14%
0.10	33	6%	47	7%	87	11%	49	13%
0.15	37	6%	47	7%	78	10%	35	9%
0.20	42	7%	67	10%	81	10%	51	13%
0.25	69	12%	61	9%	56	7%	35	9%
0.30	320	55%	303	46%	257	32%	130	33%
mae1	9	2%	19	3%	29	4%	9	2%
mae2	2	0%	11	2%	29	4%	6	2%
mar1	1	0%	3	0%	7	1%	2	1%
mar2					4	0%	1	0%
mse1	8	1%	2	0%	7	1%	3	1%
mse2	16	3%	6	1%	12	1%	7	2%
msr1	1	0%	2	0%	8	1%	2	1%
msr2	3	1%	2	0%	2	0%	5	1%

Note: For each demand pattern, the number in the parenthesis shows the total and the first and second columns show the number and percent of SKUs, respectively.

Table 6: The method for identifying initial values of statistical models

Method	# of SKUs	% of SKUs	Method	# of SKUs	% of SKUs
Erratic (582)			Lumpy (658)		
<i>Mean</i>	284	49%	<i>Mean</i>	356	54%
Naïve	258	44%	Naïve	257	39%
Optimum	40	7%	Optimum	45	7%
Intermittent (802)			Smooth (389)		
<i>Mean</i>	362	45%	Mean	150	39%
Naïve	342	43%	<i>Naïve</i>	204	52%
Optimum	98	12%	Optimum	35	9%

When the performances of the statistical forecasting methods are compared with each other (Table 7), the majority of the SKUs in each demand pattern performed best accuracies with the SBA method (61.4%; 1,493 out of 2,431 SKUs). For intermittent and lumpy demand patterns, SES (39.7%; 580 out of 1,460 SKUs) also performed better than CR.

Table 7: Performance of statistical forecasting methods according to demand patterns

Method	# of SKUs	% of SKUs	Method	# of SKUs	% of SKUs
Erratic (582)			Lumpy (658)		
CR	48	8%	CR	45	7%
SBA	446	77%	SBA	314	48%
SES	88	15%	SES	299	45%
Intermittent (802)			Smooth (389)		
CR	69	9%	CR	56	14%
SBA	452	56%	SBA	281	72%
SES	281	35%	SES	52	13%

5.4 Making judgmental forecasts

In this study, judgmental forecasts are generated by the company by using executive opinions. The company usually relies on the judgmental forecasts generated by the executives

instead of using any statistical techniques. This would create an advantage of generating accurate forecasts as they have deep experience in their business context.

For the SKUs under investigation, judgmental forecasts were generated by the purchasing specialists as experts with the best insights about the company's future. Using their experience, they decide the number of spare parts to order from their suppliers. Besides historical demand data, this study requires a proxy variable that can reproduce historical judgmental forecasts. For this purpose, 3,639 SKUs from the purchasing records covering July 2017-December 2018 (18 months) are obtained from the company. However, only 1,205 of these records could be matched with the demand data. Thus, 1,205 SKUs are used for the remaining steps of this forecasting process. This way of obtaining judgements can be accepted as equivalent to using the jury of the executive opinion method in an unstructured way to get the judgemental forecasts. It is assumed at this point that the experts made the judgmental forecasts and are aware of the variables influencing the demand for the SKUs for which they are responsible.

5.5 Combining statistical and judgmental forecasts

The next stage of the proposed framework is the combination process of the selected statistical forecasting methods that are selected for all SKUs in both datasets with the judgmental forecasts which are provided by the company.

This study aims to generate more accurate forecasts for ID by combining statistical forecasts generated by the appropriate methods through the guidance of the literature with judgmental forecasts based on expert opinions. When the time series is highly variable, as in the case of SKUs with ID patterns, it is appropriate to use judgmental forecasts. Thus, it is expected that this combination process will improve forecast accuracy. When combining the constituent forecasts for these datasets, weighted average procedures are used as this simple method performs better than the sophisticated combination strategies (Clemen, 1989; Sanders & Ritzman, 2004). Even using equal weights for the constituent forecasts (0.50 for statistical forecast and 0.50 for judgmental forecast) provides rather good performance (Clemen, 1989). Besides using simple averaging, weights ranging from 0.10 to 0.90 are used in order to see the effect of the different weights on the accuracy level. Thus, nine different combinations are implemented in the MS Excel environment.

As the judgmental forecasts provided by the company encompass 18 months, the combination procedure is applied to only this range of the dataset. A combination of statistical and judgmental forecasts is applied based on the weighted average procedure. Assume the weights for the statistical and judgmental forecasts for a SKU are 0.6 and 0.4, respectively. The combined forecast for this SKU is calculated as (Statistical Forecast Value x 0.6 + Judgmental Forecast Value x 0.4).

5.6 Evaluation

The forecast accuracy for each SKU is calculated for all of the resulting combinations. For the best-performing statistical methods, judgmental forecasts provided by the company and all combinations with different weights, MASE, RMSE, and bias are calculated for each SKU by using the "Metrics" package. The forecasting method that provides the highest accuracy is selected among the combined, statistical, and judgmental forecasts.

The performance measures for all forecasting models developed in this study are partially displayed in Table 8. It is seen that, based on three performance measures, sample SKUs

perform best in different combinations and methods. The overarching standard for selecting the best method is to look for consistent results across all measures. The criteria used in the case of inconsistent situations is to select the method with the lowest value out of the two accuracy measures (MASE and RMSE) plus the bias closest to zero.

Table 8: Performance measures for all forecast models (Partial representation)

# of SKUs	Performance Measure	S	J	0.1S 0.9J	0.2S 0.8J	0.3S 0.7J	0.4S 0.6J	0.5S 0.5J	0.6S 0.4J	0.7S 0.3J	0.8S 0.2J	0.9S 0.1J
P65	MASE	0.84	1.89	1.69	1.48	1.28	1.11	0.99	0.92	0.87	0.83	0.82
	RMSE	10.29	22.75	20.77	18.84	16.99	15.25	13.65	12.26	11.14	10.39	10.10
	Bias	2.01	20.39	18.15	15.91	13.67	11.43	9.19	6.95	4.71	2.47	0.23
P114	MASE	0.96	1.48	1.28	1.08	0.87	0.74	0.69	0.67	0.68	0.74	0.82
	RMSE	6.79	9.78	8.62	7.52	6.51	5.65	5.01	4.69	4.74	5.17	5.88
	Bias	5.41	6.83	5.61	4.39	3.16	1.94	0.71	0.51	1.74	2.96	4.18
P775	MASE	0.79	2.26	2.05	1.85	1.64	1.44	1.23	1.02	0.90	0.82	0.80
	RMSE	2.61	6.64	6.07	5.51	4.97	4.45	3.96	3.51	3.12	2.82	2.64
	Bias	0.15	6.11	5.49	4.86	4.23	3.61	2.98	2.35	1.73	1.10	0.47
P934	MASE	2.39	2.40	1.94	1.51	1.15	0.89	0.80	0.91	1.18	1.52	1.94
	RMSE	7.94	7.97	6.65	5.40	4.30	3.46	3.13	3.44	4.27	5.37	6.62
	Bias	7.29	7.33	5.87	4.41	2.95	1.48	0.02	1.44	2.91	4.37	5.38

Note: S stands for Statistical Model and J stands for Judgmental Forecast.

Table 9 lists how many SKUs were chosen as the top performers in the corresponding settings. These results show that the majority (64.2%; 772 out of 1,205) of SKUs did perform better when judgmental and statistical forecasts were combined. However, for many SKUs, the combinations that give statistical forecasts more weight performed higher. For the pure statistical models, the best performance was observed for 33.4% (402 out of 1,205) of SKUs. On the other hand, the judgmental forecasts provided by the company’s purchasing specialists performed worse (2.4%) compared to the alternative approaches. Models with statistical forecasts given a weight greater than or equal to 0.5 performed better for 666 SKUs (55.3%). To sum up, most of the most accurate forecasts are generated by using the combination process.

Table 9: Overall comparison of all models based on p and CV^2

Model	# of SKUs	Average p	Average CV^2
0.1S - 0.9J	10	1.76	0.81
0.2S - 0.8J	19	2.19	0.69
0.3S - 0.7J	38	1.79	0.98
0.4S - 0.6J	41	1.89	0.71
0.5S - 0.5J	120	1.87	0.73
0.6S - 0.4J	68	1.36	0.84
0.7S - 0.3J	109	1.37	0.74
0.8S - 0.2J	162	1.30	0.74
0.9S - 0.1J	207	1.27	0.58
Judgmental	29	2.44	0.79
Statistical	402	1.44	0.56
Average	1,205	14.88	0.66

For getting additional insights into the results, the average p and CV^2 values in each sub-group are also computed. To investigate the relationship between the weights assigned to statistical forecasts and the parameters describing ID characteristics (i.e., average inter-demand interval, p and CV^2 values), an OLS regression analysis is performed.

Analysis showed that 70% of the changes in the weights assigned to statistical forecasts could be explained by the average inter-demand interval p and CV^2 values ($F=29.08$ with p -value <0.05). The relationship between the weights assigned to statistical forecasts and CV^2 values is found as negative ($t=-3.136$ with p -value <0.05). The negative relationship between the weights assigned to statistical forecasts and the value of p is also substantiated ($t= -5.98$ with p -value <0.05). For higher inter-demand intervals and CV^2 values, lower weights may be preferred for statistical forecasts in combined models. However, this finding needs to be tested with stronger evidence in another research specifically designed for exploring such connections.

6 Conclusion

Combining forecasts in the context of ID is one of the unnoticed research areas in the literature. Studies combining statistical and judgmental forecasts for the ID of SKUs are particularly scarce. From these facts, this study developed a methodological framework for combining statistical and judgmental forecasts for ID. The proposed framework for forecasting is described as a six-stage model to obtain more accurate results for SKUs with ID patterns. In line with the simple definition of [Green & Armstrong \(2015\)](#), the study defined this framework as processes that are understandable to forecast users. Then, the framework is tested using real-world data that includes monthly customer demands for after-sales spare parts from an anonymous company that manufactures small home appliances in the electronics industry. Results showed that combination is the best method for the majority of SKUs, as expected.

This paper contributes to the limited literature by addressing the gap between the combined forecast and the ID forecast. Companies primarily employ judgmental approaches in which their forecasts are entirely dependent on business context information. Thus, these kinds of pure judgmental forecasts have the risk of being biased. The proposed framework gives practitioners and researchers a comprehensive overview to help them make more accurate forecasts while also encouraging the use of simple but structured approaches.

On the other hand, the results of this study must be interpreted in light of some limitations. The first is concerned with data quality and missing information. The company provided the data without giving any details for the SKUs, such as what the SKU is, if SKU is a new item or not used anymore, and whether it has a similarity to other SKUs or not. Thus, all calculations are done assuming that the SKUs are independent of each other and are actively used for the given periods. More detail about the SKUs would help to categorize and arrange the dataset for the forecasting process, which would provide better accuracy levels. The second limitation is the method used to generate judgmental forecasts. During the implementation, company representatives generate judgmental forecasts as is standard procedure in their operations. At this point, it is assumed that the representatives made the judgmental forecasts as experts and are aware of the variables influencing demand for the SKUs for which they are responsible. To better measure, the model's performance, this part of the implementation could use more structured and controlled methods, as suggested in this study.

Future research can determine the framework's applicability in different contexts. When developed, a toolkit can help and motivate people to put this model into action.

References

- Armstrong, J. S. (1985). *Long Range Forecasting: From Crystal Ball to Computer* (2nd ed.). Wiley-Interscience, New York.
- Armstrong, J. S. (2001a). Combining Forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Dordrecht. https://repository.upenn.edu/marketing_papers/150.
- Armstrong, J. S. (2001b). Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Dordrecht. https://repository.upenn.edu/marketing_papers/150.
- Armstrong, J. S. (2001c). Selecting Forecasting Methods. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Dordrecht. https://repository.upenn.edu/marketing_papers/150.
- Babai, M. Z., Dallery, Y., Boubaker, S., & Kalai, R. (2019). A New Method to Forecast Intermittent Demand in the Presence of Inventory Obsolescence. *International Journal of Production Economics*, 209, 30-41. doi:10.1016/j.ijpe.2018.01.026
- Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, 20(4), 451-468. doi:10.1057/jors.1969.103
- Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerial Intuition: 50% Model + 50% Manager. *Management Science*, 36(8), 887-899. doi:10.1287/mnsc.36.8.887
- Boylan, J. E., Syntetos, A. A., & Karakostas, G. C. (2008). Classification for Forecasting and Stock Control: A Case Study. *Journal of the Operational Research Society*, 59(4), 473-481. doi:10.1057/palgrave.jors.2602312
- Bunn, D. W., & Salo, A. A. (1993). Forecasting with Scenarios. *Journal of Operational Research*, 68(3), 291-303. doi:10.1016/0377-2217(93)90186-Q
- Chan, F., & Pauwels, L. L. (2018). Some Theoretical Results on Forecast Combinations. *International Journal of Forecasting*, 34(1), 64-74. doi:10.1016/j.ijforecast.2017.08.005
- Claeskens, G., Magnus, J. R., Vasnev, A., & Wang, W. (2016). The Forecast Combination Puzzle: A Simple Theoretical Explanation. *International Journal of Forecasting*, 32(3), 754-762. doi:10.1016/j.ijforecast.2015.12.005
- Clemen, R. T. (1989). Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting*, 5(4), 559-583. doi:10.1016/0169-2070(89)90012-5
- Constantino, F., Di Gravio, G., Patriarca, R., & Petrella, L. (2018). Spare parts management for irregular demand items. *Omega*, 81, 57-66. doi:10.1016/j.omega.2017.09.009
- Croston, J. D. (1972). Forecasting and Stock Control for Intermittent Demands. *Operational Research Quarterly (1970-1977)*, 23(3), 289-303. doi:10.2307/3007885
- Dalkey, N. (1969). An Experimental Study of Group Opinion: The Delphi Method. *Futures*, 1(5), 408-426. doi:10.1016/S0016-3287(69)80025-X
- Davydenko, A., & Fildes, R. (2013). Measuring Forecasting Accuracy: The Case of Judgmental Adjustments to SKU-level Demand Forecasts. *International Journal of Forecasting*, 29(3), 510-522. doi:10.1016/j.ijforecast.2012.09.002

- Duncan, G., Gorr, W. L., & Szczypula, J. (2001). Forecasting analogous time series. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Dordrecht.
- Eaves, A. H. (2002). *Forecasting for the Ordering and Stock-holding of Consumable Spare Parts* (PhD thesis). Department of Management Science, The Management School, Lancaster University.
- Eaves, A. H., & Kingsman, B. G. (2004). Forecasting for the Ordering and Stock-holding of Consumable Spare Parts. *Journal of the Operational Research Society*, 55(4), 431-437. doi:10.1057/palgrave.jors.2601697
- Eroğlu, C., & Croxton, K. L. (2010). Biases in Judgmental Adjustments of Statistical Forecasts: The Role of Individual Differences. *Journal of the Operational Research Society*, 26(1), 116-133. doi:10.1016/j.ijforecast.2009.02.005
- Fildes, R. (1992). The Evaluation of Extrapolative Forecasting Methods. *International Journal of Forecasting*, 8(1), 81-98. doi:10.1016/0169-2070(92)90009-X
- Fildes, R., Ma, S., & Kolassa, S. (2019). *Retail Forecasting: Research and Practice* (MPRA Working Paper No. 89356). Munich Archive. https://mpra.ub.uni-muenchen.de/89356/1/MPRA_paper.89356.pdf.
- Fildes, R., & Petropoulos, F. (2015). Simple versus Complex Selection Rules for Forecasting Many Time Series. *Journal of Business Research*, 68(8), 1692-1701. doi:10.1016/j.jbusres.2015.03.028
- Fischer, I., & Harvey, N. (1999). Combining Forecasts: What Information do Judges Need to Outperform the Simple Average? *International Journal of Forecasting*, 3(15), 227-246. doi:10.1016/S0169-2070(98)00073-9
- Franses, P. H., & Legerstee, R. (2010). Do Experts' Adjustments on Model-Based SKU-Level Forecasts Improve Forecast Quality? *Journal of Forecasting*, 29(3), 331-340. doi:10.1002/for.1129
- Ghobbar, A. A., & Friend, C. H. (2002). Sources of Intermittent Demand for Aircraft Spare Parts within Airline Operations. *Journal of Air Transport Management*, 8(4), 221-231. doi:10.1016/S0969-6997(01)00054-0
- Ghobbar, A. A., & Friend, C. H. (2003). Evaluation of Forecasting Methods for Intermittent Parts Demand in the Field of Aviation: A Predictive Model. *Computers & Operations Research*, 30(14), 2097-2114. doi:10.1016/S0305-0548(02)00125-9
- Green, K. C., & Armstrong, J. S. (2007). Structured Analogies for Forecasting. *International Journal of Forecasting*, 23(3), 365-376. doi:10.1016/j.ijforecast.2007.05.005
- Green, K. C., & Armstrong, J. S. (2015). Simple Versus Complex Forecasting: The Evidence. *Journal of Business Research*, 68(8), 1678-1685. doi:10.1016/j.jbusres.2015.03.026
- Gutierrez, R. S., Solis, A. O., & Mukhopadhyay, S. (2008). Lumpy Demand Forecasting using Neural Networks. *International Journal Production Economics*, 111(2), 409-420. doi:10.1016/j.ijpe.2007.01.007
- Hanke, J. E., & Wichern, D. (2014). *Business Forecasting* (9th ed.). Pearson Education Limited, Essex.
- Hasni, M., Babai, M. Z., Aguir, M. S., & Jemai, Z. (2019). An Investigation on Bootstrapping Forecasting Methods for Intermittent Demands. *International Journal Production Economics*, 209, 20-29. doi:10.1016/j.ijpe.2018.03.001

- Hsu, C., & Sandford, B. (2007). The Delphi Technique: Making Sense of Consensus. *Practical Assessment, Research, and Evaluation*, 12(10), 1-8. doi:10.7275/pdz9-th90
- Hua, Z. S., Zhang, B., Yang, J., & Tan, D. S. (2007). A New Approach of Forecasting Intermittent Demand for Spare Parts Inventories in the Process Industries. *Journal of the Operational Research Society*, 58(1), 52-61. doi:10.1057/palgrave.jors.2602119
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (Second ed.). Otexts, Melbourne.
- Hyndman, R. J., & Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 22(4), 679-688. doi:10.1016/j.ijforecast.2006.03.001
- Johnston, F. R., Boylan, J., & Shale, E. (2003). An Examination of the Size of Orders from Customers, their Characterisation and the Implications for Inventory Control for Slow-moving Items. *The Journal of the Operational Research Society*, 54(8), 833-837. doi:10.1057/palgrave.jors.2601586
- Johnston, F. R., & Boylan, J. E. (1996). Forecasting for Items with Intermittent Demand. *The Journal of the Operational Research Society*, 47(1), 113-121. doi:10.2307/2584256
- Kostenko, A. V., & Hyndman, R. J. (2006). A Note on the Categorization of Demand Patterns. *The Journal of the Operational Research Society*, 57, 1256-1257. doi:10.1057/palgrave.jors.2602211
- Kourentzes, N. (2013). Intermittent Demand Forecasts with Neural Networks. *International Journal of Production Economics*, 143(1), 198-206. doi:10.1016/j.ijpe.2013.01.009
- Kourentzes, N. (2014). On Intermittent Demand Model Optimisation and Selection. *International Journal of Production Economics*, 156(1), 180-190. doi:10.1016/j.ijpe.2014.06.007
- Kourentzes, N., & Petropoulos, F. (2016). *R: Package 'tsintermittent'* [R package manual]. <https://cran.r-project.org/web/packages/tsintermittent/tsintermittent.pdf>.
- Lawrence, M., Goodwin, P., O'Connor, M., & Önköl, D. (2006). Judgmental Forecasting: A Review of Progress over the Last 25 Years. *International Journal of Forecasting*, 22(3), 493-518. doi:10.1016/j.ijforecast.2006.03.007
- Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1986). The Accuracy of Combining Judgmental and Statistical Forecasts. *Management Science*, 32(12), 1521-1532. doi:10.1287/mnsc.32.12.1521
- Levén, E., & Segerstedt, A. (2004). Inventory Control with a Modified Croston Procedure and Erlang Distribution. *International Journal of Production Economics*, 90(3), 361-367. doi:10.1016/S0925-5273(03)00053-7
- Li, L., Kang, Y., Petropoulos, F., & Li, F. (2022). Feature-based Intermittent Demand Forecast Combinations: Accuracy and Inventory Implications. *International Journal of Production Research*. doi:10.1080/00207543.2022.2153941
- Makridakis, S. (1993). Accuracy Measures: Theoretical and Practical Concerns. *International Journal of Forecasting*, 9(4), 527-529. doi:10.1016/0169-2070(93)90079-3
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 Competition: Background, Organization, and Implementation. *International Journal of Forecasting*, 38(4), 1325-1336. doi:10.1016/j.ijforecast.2021.07.007

- Nagaria, P. (2017). *Forecasting intermittent demand: Traditional smoothing approaches versus the Croston method*. https://www.oreilly.com/library/view/strata-data-conference/9781491985373/video317307.html?_gl=1*195ftmp*_ga*NTQxODkwNjQyLjE2NzQyNzEyOTI.*_ga_092EL089CH*MTY3NDI3MTI5MS4xLjEuMTY3NDI3MTg2OC41MC4wLjA.
- Palm, F. C., & Zellner, A. (1992). To Combine or not to Combine? Issues of Combining Forecasts. *Journal of Forecasting*, 11(8), 687-701. doi:10.1002/for.3980110806
- Parackal, M., Goodwin, P., & O'Connor, M. (2007). Judgement in Forecasting (Editorial). *International Journal of Forecasting*, 23(3), 343-345. doi:10.1016/j.ijforecast.2007.05.004
- Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do 'Big Losses' in Judgmental Adjustments to Statistical Forecasts Affect Experts' Behaviour? *European Journal of Operational Research*, 249(3), 842-852. doi:10.1016/j.ejor.2015.06.002
- Petropoulos, F., & Kourentzes, N. (2015). Forecast Combinations for Intermittent Demand. *Journal of the Operational Research Society*, 66(6), 914-924. doi:10.1057/jors.2014.62
- Petropoulos, F., & Svetunkov, I. (2020). A Simple Combination of Univariate Models. *International Journal of Forecasting*, 36(1), 110-115. doi:10.1016/j.ijforecast.2019.01.006
- Piñe, Ç., Turrini, L., & Meissner, J. (2021). Intermittent Demand Forecasting for Spare Parts: A Critical Review. *Omega*, 105, 1-30. doi:10.1016/j.omega.2021.102513
- Pour, A. N., Tabar, B. R., & Rahimzadeh, A. (2008). A Hybrid Neural Network and Traditional Approach for Forecasting Lumpy Demand. *International Journal of Industrial and Manufacturing Engineering*, 2(4), 1028-1034. doi:10.5281/zenodo.1075923
- Qian, W., Rolling, C., Cheng, G., & Yang, Y. (2019). On the Forecast Combination Puzzle. *Econometrics*, 7(39), 1-26. doi:10.3390/econometrics7030039
- Saccani, N., Visintin, F., Mansini, R., & Colombi, M. (2017). Improving Spare Parts Management for Field Services: A Model and a Case Study for the Repair Kit Problem. *IMA Journal of Management Mathematics*, 28(2), 185-204. doi:10.1093/imaman/dpw023
- Sanders, N. R. (1992). Accuracy of Judgmental Forecasts: A Comparison. *Omega*, 20(3), 353-364. doi:10.1016/0305-0483(92)90040-E
- Sanders, N. R., & Manrodt, K. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega*, 31(6), 511-522. doi:10.1016/j.omega.2003.08.007
- Sanders, N. R., & Ritzman, L. P. (1992). The Need for Contextual and Technical Knowledge in Judgmental Forecasting. *Journal of Behavioral Decision Making*, 5(1), 39-52. doi:10.1002/bdm.3960050106
- Sanders, N. R., & Ritzman, L. P. (2001). Judgmental Adjustment of Statistical Forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Dordrecht.
- Sanders, N. R., & Ritzman, L. P. (2004). Integrating judgmental and quantitative forecasts: methodologies for pooling marketing and operations information. *International Journal of Operations & Production Management*, 24(5), 514-529. doi:10.1108/01443570410532560
- Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory management and production planning and scheduling* (3rd ed.). John Wiley & Sons, New York.

- Smith, J., & Wallis, K. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3), 331-355. doi:10.1111/j.1468-0084.2008.00541.x
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books, New York.
- Syntetos, A. A. (2001). *Forecasting of intermittent demand* (PhD thesis). Business School Buckinghamshire Chilterns University College, Brunel University, UK.
- Syntetos, A. A., Babai, M. Z., & Gardner Jr., E. S. (2015). Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. *International Journal of Production Economics*, 68(8), 1746–1752. doi:10.1016/j.jbusres.2015.03.034
- Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics*, 71(1-3), 457-466. doi:10.1016/S0925-5273(00)00143-2
- 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. doi:10.1057/palgrave.jors.2601841
- Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics*, 118(1), 72–81. doi:10.1016/j.ijpe.2008.08.011
- Teunter, R., & Duncan, L. (2009). Forecasting intermittent demand: a comparative study. *Journal of the Operational Research Society*, 60(3), 321–329. doi:10.1057/palgrave.jors.2602569
- Timmermann, A. (2006). Forecast Combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, p. 135–196). Elsevier. doi:10.1016/S1574-0706(05)01004-9
- van Kampen, T. J., Akkerman, R., & van Donk, D. P. (2012). SKU classification: a literature review and conceptual framework. *International Journal of Operations & Production Management*, 32(7), 850–876. doi:10.1108/01443571211250112
- Waller, D. (2015). *Methods for Intermittent Demand Forecasting* (Working Paper). Lancaster University. http://www.lancaster.ac.uk/pg/waller/pdfs/Intermittent_Demand_Forecasting.pdf.
- Wallström, P., & Segerstedt, A. (2010). Evaluation of forecasting error measurements and techniques for intermittent demand. *International Journal of Production Economics*, 128(2), 625–636. doi:10.1016/j.ijpe.2010.07.013
- Wang, X., Hyndman, R., Li, F., & Kang, Y. (2022). *Forecast Combinations: An over 50-year Review* (Preprint). arXiv. <https://arxiv.org/pdf/2205.04216>.
- Weinberg, C. B. (1986). Arts Plan: Implementation, Evolution, and Usage. *Marketing Science*, 5(2), 143–158. doi:10.1287/mksc.5.2.143
- Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, 20(3), 375–387. doi:10.1016/S0169-2070(03)00013-X
- Willemain, T. R., Smart, C. N., Shockor, J. H., & DeSautels, P. A. (1994). Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. *International Journal of Forecasting*, 10(4), 529–538. doi:10.1016/0169-2070(94)90021-3

- Williams, T. M. (1984). Stock control with sporadic and slow-moving demand. *The Journal of the Operational Research Society*, 35(10), 939–948. doi:10.1057/jors.1984.185
- Wilson, J. H., & Keating, B. (2008). Introduction to Business Forecasting. In J. H. Wilson & B. Keating (Eds.), *Business Forecasting with ForecastX* (p. 1-55). McGraw Hill-Irwin, New York.
- Wright, T. C. (2013). *Real Life Examples of Qualitative Forecasting*. <https://smallbusiness.chron.com/real-life-examples-qualitative-forecasting-72990.html>.

Appendix: Details of performance measures

Measure	Formula	Explanation
<i>Scale-dependent Error Measures</i>		
Mean Error	$ME = \frac{1}{n} \sum_{t=1}^n e_t$	The small value of ME does not mean that the error is small but shows biases. ME takes a negative (positive) value if the forecasting method underpredicts (overpredicts) the actual.
Cumulative Error	$CE = \sum_{t=1}^n e_t$	This method sums up the errors to determine the bias.
Periods in stock	$PIS_i = -\sum_{h=1}^H \sum_{j=1}^h (Y_{N+j} - \hat{Y}_j)$	H is the length of the required forecast horizon. This measure shows the total number of periods in stock or stock-out periods of the forecasted item. Positive PIS indicates stock left over, whereas negative PIS represents stock shortages.
Mean Absolute Error	$MAE = \frac{1}{n} \sum_{t=1}^n e_t $	MAE shows the errors regardless of under or over the forecast. This method gives an average of error measurement irrespective of the direction. Also, it eliminates the canceling of the problem caused by the ME method. This method shows how large the average error can be.
Mean Square Error	$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2$	Like MAE, it eliminates the canceling out problem. This penalizes large forecasting errors by squaring the errors (Hanke & Wichern, 2014). MSE is widely used to compare the accuracy levels of different methods. MSE value close to 0 is preferable
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$	Lower values of RMSE is better.

Scale-independent Error Measures		
Mean absolute scaled error	$MASE = \frac{1}{n} \sum_{t=1}^n \left \frac{e_t}{\frac{1}{(n-1)} \sum_{t=2}^n A_t - A_{t-1} } \right $	Its value less (higher) than 1 indicates that the actual forecast performance is better (worse) than the naïve method. Division by zero only occurs if all the values in the time series are equal. It is symmetrical and robust to outliers.
Scaled CE	$sCE = \frac{\sum Y_{N+h} - F_h}{\frac{1}{N} \sum_{t=1}^N Y_t}$	N is the number of in-sample observations, Y_{N+h} is the h^{th} out-of-sample period and F_h is the h-steps ahead forecasts. sCE can be used for comparing the performance of different methods on different datasets.
Scaled PIS	$sPIS_i = \frac{PIS_i}{\frac{1}{N} \sum_{t=1}^N Y_t}$	A scale-independent form of PIS.
Scaled MAE	$sAE_{i,h} = \frac{ Y_{N+h} - F_h }{\frac{1}{N} \sum_{t=1}^N Y_t}$	Like in sCE, MAE can be scaled to be able to average the measures across series. sMAE is the scaled absolute error averaged over all series and horizons.
Scaled MSE	$sSE_{i,h} = \left(\frac{Y_{N+h} - F_h}{\frac{1}{N} \sum_{t=1}^N Y_t} \right)^2$	sMSE (Scaled MSE) is the mean of scaled squared error across all series and horizons.
Root Mean Squared Scaled Error	$RMSSE = \sqrt{\frac{1}{h} \frac{\sum_{t=n+1}^{(n+h)} (Y_t - \hat{Y}_t)^2}{\frac{1}{1-n} \sum_{t=2}^n (Y_t - Y_{t-1})^2}}$	RMSSE is a variant of the well-known MASE.
Percentage Error Measures		
Mean Absolute Percentage Error	$MAPE = \frac{1}{n} \sum_{t+1}^n \left \frac{e_t}{A_t} \right \times 100$	If there are zero values in the denominator (actual demand), there will be division by zero problems. Thus, the percentage calculation will be undefined. Because of this, using MAPE is not appropriate for items with ID.
Symmetric Mean Absolute Percentage Error	$S - MAPE = \frac{2}{n} \sum_{t+1}^n \frac{ A_t - F_t }{A_t + F_t} \times 100$	This method is proposed by Makridakis (1993) to avoid asymmetry caused by the application of MAPE. Unlike MAPE, S-MAPE has a lower and an upper bound. The formula above provides results between 0 and 200%.

Relative Error Measures		
Relative Geometric Root Mean Square Error	$RGRMSE = \frac{(\prod_{t+1}^n (A_t - F_{B,t})^2)^{1/2n}}{(\prod_{t+1}^n (A_t - F_{A,t})^2)^{1/2n}}$	Subscripts A and B refers to the forecasting methods. RGRMSE is a safe measure to use in ID.
Percentage of times Better (PB)	PB is a method under consideration that outperforms another.	It can easily be calculated and interpreted and is robust to large forecast errors. PB is used for comparing two estimators based on performance criteria like mean deviation or geometric mean square error.
Percentage Best (PB_t)	PB_t is the percentage of occurrences when one method outperforms all others.	When more than two estimators are compared, Percentage Best (PB_t) is used rather than PB.