





# Comparison of selection and combination strategies for demand forecasting methods

Saymon Galvão Bandeira<sup>a\*</sup> , Symone Gomes Soares Alcalá<sup>b</sup> , Roberto Oliveira Vita<sup>c</sup> ,  
Talles Marcelo Gonçalves de Andrade Barbosa<sup>a</sup> 

<sup>a</sup>Pontifícia Universidade Católica de Goiás, Goiânia, GO, Brasil

<sup>b</sup>Universidade Federal de Goiás, Faculdade de Ciências e Tecnologia, Campus Aparecida de Goiânia-Regional Goiânia, Aparecida de Goiânia, GO, Brasil

<sup>c</sup>Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Porto, Portugal

\*saymongb@gmail.com, saymongb@hotmail.com

## Abstract

**Paper aims:** In this study, effective strategies to combine and select forecasting methods are proposed. In the selection strategy, the best performing forecasting method from a pool of methods is selected based on its accuracy, whereas the combination strategies are based on the mean methods' outputs and on the methods' accuracy.

**Originality:** Despite the large amount of work in this area, the actual literature lacks of selection and combination strategies of forecasting methods for dealing with intermittent time series.

**Research method:** The included forecasting methods are state-of-the-art approaches applied to industrial and academics forecasting problems. Experiments were performed to evaluate the performance of the proposed strategies using a spare part data set of an industry of elevators and a data set from the M3-Competition.

**Main findings:** The results show that, in most cases, the accuracy of the demand forecasts can be improved when using the proposed selection and combination strategies.

**Implications for theory and practice:** The proposed methodology can be applied to forecasting problems, covering a variety of characteristics (e.g., intermittency, trend). The results reveal that combination strategies have potential application, perform better than state-of-the-art models, and have comparable accuracy in intermittent series. Thus, they can be employed to improve production planning activities.

## Keywords

Time series forecasting. Forecast uncertainty. Technology forecasting. Combination strategies. Forecasting method selection.

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

Received: Jan. 31, 2020; Accepted: Aug. 10, 2020.

## 1. Introduction

The introduction of digital technologies on the industry to provide integration between physical and digital systems has emerged under the form of Industry 4.0 (Frank et. al., 2019). These technologies can provide useful data to manufacturing systems. In particular, in smart manufacturing systems, planning activities can rely on information provided by intelligent computer systems, which use historical data to generate valuable information (e.g. future product demands).

Demand forecast refers to predict or estimate the need for a product or a component in a future time period (Armstrong, 2001). The information about the forecasted demand can be used to support managerial decisions



and planning process activities in operations. For example, in Guo et al. (2017), the forecasts are employed to support ordering decisions of airplane spare parts. Yu et al. (2011) proposes forecasting models to estimate demands of fashion products. Another example of application is presented by Syntetos et al. (2005), where the main objective is to forecast spare parts demands from an automotive industry.

The accuracy of the demand forecasts is important for companies, since forecasts are usually employed as input to inventory systems (Wang & Petropoulos, 2016; Rego & Mesquita, 2015; Babai et al., 2019). Several demand forecasting methods have been proposed in the literature, such as, Simple Exponential Smoothing (SES) (Hyndman & Athanasopoulos, 2018), Croston (CR) (Croston, 1972), among others; and many studies have been devoted to the selection of the appropriate forecasting method, which can depend on the time series characteristics, performance or professional expertise (Syntetos et al., 2005; Petropoulos et al., 2018; Moon et al., 2013).

In literature, different selection criteria for forecasting models have been used. The selection can be based on the time series characteristics (Syntetos et al., 2005; Petropoulos et al., 2018; Heinecke et al., 2013), on the forecasting model performance (Wang & Petropoulos, 2016; Fildes & Petropoulos, 2015), on the information criteria (Qi & Zhang, 2001), or on the judgmental expert selection (Petropoulos et al., 2018). For example, Adya et al. (2001) proposed an automated framework to identify six different time series characteristics in a rule-based forecast system.

Rule-based forecast is an expert system proposed by Collopy & Armstrong (1992), which relies on 28 characteristics of time series to weight four forecasting methods. In Adya et al. (2001), another strategy is proposed. The presence of outliers, level shifts, changes in trend, unstable recent trend, functional form, and unusual (last) demands are considered as time series characteristics. Another example is the selection based on an information criteria proposed by Qi & Zhang (2001). Using financial time series from S&P 500 Index, a strategy selects among many Artificial Neural Network and Autoregressive (AR) models using the Akaike information criteria and the Bayesian information criteria.

Furthermore, combinations of forecasting methods can significantly improve the accuracy of forecasts and reduce the variance of prediction errors, which are desirable characteristics for inventory purposes (Wang & Petropoulos, 2016). Combination of forecasts is a process of using different forecasting models to produce a final forecast. The application of combination schemes avoids the implicit assumptions about the underlying process of data generation. Kourentzes et al. (2019) proposed a heuristic, based on quartiles definition using forecasting errors, to build a pool of forecasting models for combination and selection. The approach was evaluated using the M3-Competition data; however, the employed evaluation metrics are different from those employed to evaluate the M3-Competition results, which difficult the approach evaluation and comparison. Barrow & Kourentzes (2016) analyze the impact of a forecasting combination in terms of forecast error distribution and safety stock using demand data of a consumer goods manufacturer. The authors revealed that forecasts from a combination of Naive Forecast (NF), SES, AR, Autoregressive Integrated Moving Average (ARIMA), Theta and Multiple Aggregation Prediction Algorithm (MAPA) models can improve inventory decisions (Barrow & Kourentzes, 2016). However, the work does not compare different combination strategies.

In addition, Guo et al. (2017) evaluated a double-level combination of forecasting methods to predict spare part data of an aircraft fleet. The work employed different data types (e.g., flight time, number of takeoffs, number of landings, among others) that influence the spare part consumption to design the forecasting methods. The used methods are Exponential Smoothing methods variations, Genetic Neural Networks and Grey model. The proposed combination strategy consists of assigning the weights each forecasting method by solving a quadratic programming problem (“low-level combination”) and using a genetic algorithm for a “top-level combination”. The proposed approach outperformed other forecasting models. However, high computational time is required to produce predictions due to the use of a neural network that requires several computations to determine optimal weights.

Also, Wang & Petropoulos (2016) employed a simple 50-50% combination of two forecasting sources, and results have shown improvements in the accuracy forecasts. The proposed combination consists of a simpler scheme, where the assigned weights are equals for two forecasting sources: a forecasting model from commercial software, and forecasts judgmentally produced by experts. The strategies were compared to the other combinations as the selection based on the variance of the forecast error, the selection based on the Mean Absolute Error (MAE), and a single forecasting method. The proposed combination strategies minimized the total cost of inventory meanwhile maximized the forecast accuracy, however, the strategies were evaluated only on a stationary data from a pharmaceutical industry.

On the other hand, Franses & Legerstee (2011) propose the use of a combination of forecasting methods with experts forecast. The proposed combination strategy also considers the specific characteristics of experts to assign

optimal weights. The results of the evaluation revealed that the 50-50% combination, using a statistical model and an expert model, can be a useful strategy. However, the strategy can hide the contribution of each model on the final forecast, since it can assign zero weight for one of them; although, the literature suggests that the use of both is usually better. Therefore, several works have demonstrated the predictive accuracy of combinations in forecasting problems.

Nonetheless, most studies do not deal with intermittent data series (i.e. series with a large number of zero values), which are common in the spare part industry. Intermittent demand time series can lead to high costs of holding due to the high risk of obsolescence (Babai et al., 2019). This type of time series is especially difficult to forecast, because they are limited to non-zero demand data and have high variability of values. In stock control systems, inappropriate levels can lead to out of stock or excessive quantities for intermittent items. Specific methods have been proposed (Croston, 1972; Syntetos et al., 2005; Babai et al., 2019). A popular method for this type of time series was proposed by Croston (1972), and it achieves empirically good performance in many studies. Although, some works have pointed out the existence of positive bias in this method. For example, Syntetos et al. (2005) evaluated the performance of intermittent and traditional forecasting methods using data from spare parts from an automotive industry. In this case, an adjusted version of the Croston's method, called Syntetos and Boylan Approximation (SBA), achieved superior performance than the Croston's method.

Moreover, Babai et al. (2019) propose and evaluate an approach for intermittent demands, called modified SBA method, using data from the military sector and the automotive industry. The approach is efficient to forecast intermittent time series, since traditional forecasting approaches do not make appropriate adjustments when no demand occurs. However, the study lacks of comparisons between the proposed approach and other approaches to handle intermittent data.

To address the listed problems, this paper proposes and compares one forecasting method selection strategy and two forecasting method combination strategies for demand forecasting problems with different characteristics (e.g. intermittency, trend, stationary and nonstationary). In the forecasting method selection strategy, the best forecasting method from a pool of forecasting methods is selected based on its accuracy on a validation interval. On the other hand, the forecasting method combination strategies are based on the mean methods' outputs and on the methods' accuracy. Forecasting method combinations are employed in this work because researches have demonstrated that they improve the generalization capability and overall performance of the systems (Soares et al., 2012); and thus, they have advantage over a single individual forecasting model in terms of forecasting accuracy (Choi & Lee, 2018).

The strategies are developed using SES, Holt's linear trend method (HOLT) (a variant of SES), CR, AR and NF as forecasting models. The main contribution of this work is to propose a set of forecasting models with heterogeneous capabilities, so that the proposed strategies can achieve good results on time series with different data characteristics (such as, intermittency, increasing or decreasing patterns, stationary and nonstationary). Experiments, using a spare part data set (with intermittent demand) from an industry of elevators and a data set from the M3-Competition (Makridakis & Hibon, 2000), are reported to demonstrate the performance of the proposed strategies. The main contributions of this paper are to propose and compare a number of selection and combination strategies for intermittent and non-intermittent time series.

The rest of this paper is organized as follows. Section 2 describes the proposed combination strategy selection procedures, and evaluation metrics. Moreover, it presents the proposed forecasting method strategies. Section 3 presents and discusses the main results of this paper. In Section 4, concluding remarks of this paper are summarized.

## 2. Proposed forecasting approaches

This section describes the proposed approaches for demand forecasting. It starts describing the main concepts about forecasting methods and the employed forecasting methods in this paper (Subsection 2.1). Then, Subsection 2.2 details the main evaluation metrics for forecasting methods. Subsection 2.3 introduces combination strategies for forecasting methods, and presents the proposed combination strategies in this paper. Finally, Subsection 2.4 presents the main strategies for forecasting method selection, and describes the proposed forecasting method selection in this work. Table 1 describes the main employed nomenclature in this paper.

### 2.1. Forecasting methods

In this paper, the demand forecasting methods were selected to cover a wide range of characteristics in time series (such as, trend, intermittency, autocorrelation, among others). Moreover, state-of-the-art methods in literature and employed methods on commercial software were selected in this research. As described previously, the selected methods are SES, HOLT, CR, AR and NF. Table 2 presents the main notations of this work.

Table 1. Nomenclature.

Term	Description
AR( <i>p</i> )	Autoregressive model of order <i>p</i>
AUTO	Automatic selection of forecasts based on the accuracy
CF <sub>m</sub>	Combination Forecast based on the mean
CF <sub>w</sub>	Combination Forecast based on the weighted mean
CR	Croston method
HOLT	A Holt's linear method based on SES
MASE	Mean Absolute Scaled Error
NF	Naive Forecast
RMSE	Root Mean Square Error
SES	Simple Exponential Smoothing

Table 2. Notations.

Symbol	Meaning
$e_t$	A forecast error, $y_t - \hat{y}_t$ , computed on time $t$ .
$y_t$	Observed demand on time $t$ .
$h$	Size of the forecast horizon.
$\hat{y}_{t+h}$	A forecast on time $t + h$ .
$z_t$	Demand size forecast for time period after $t$ .
$p_t$	Demand interval between the demand in period $t$ and the previous demand.
$\hat{p}$	Demand interval forecast for time period after $t$ .
$w$	Vector of weights in the combination of forecasts.
$o$	The forecast output vector of methods.
$N$	Number of observations

NF is a well-known forecasting method, widely used in literature as a benchmark for performing comparisons among other forecasting methods (Franses & Legerstee, 2011; Fildes & Petropoulos, 2015; Wang & Petropoulos, 2016). This method assumes that the last observation in data time series is the most important data. In this case, an obtained estimate by the NF method is equal to the last observed demand on the data, that is:

$$\hat{y}_{t+h} = y_t \tag{1}$$

in Equation 1,  $y_t$  is the observed (real) demand on time  $t$  and  $\hat{y}_{t+h}$  is a forecast (prediction) on time  $t + h$ .

The SES method is also a popular forecasting method and it can be found in several commercial software, such as, SAP, Oracle RDF and ForecastPro. SES is usually employed when there is no clear pattern of trend or seasonality on a time series (Hyndman & Athanasopoulos, 2018). A forecast made by the SES method is determined as:

$$\hat{y}_{t+h} = \alpha y_t + (1 - \alpha) \hat{y}_t \tag{2}$$

in Equation 2,  $0 \leq \alpha \leq 1$  is a smoothing parameter. The SES method works by weighting past observations, where the weights decrease exponentially over time as the observations get older. To deal with time series with trends, a variant of the SES method, called HOLT, was included in this work. The Holt's linear trend method is a modification of SES on which the forecast value is decomposed into level and trend components, being the trend component is calculated using  $h$  (Holt, 2004). A prediction using the HOLT method can be performed as in Equation 3:

$$\hat{y}_{t+h} = l_t + hb_t \tag{3}$$

$$l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \tag{4}$$

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \tag{5}$$

in Equation 4 and Equation 5,  $0 \leq \alpha \leq 1$  and  $0 \leq \beta \leq 1$  are the smoothing parameters,  $l_t$  is a forecast of the level of the series at time  $t$ , and  $b_t$  is a forecast of the trend (slope) of the series at time  $t$ .

The literature proposes specific methods to deal with intermittent demand time series (Croston, 1972; Syntetos et al., 2005), which are characterized by multiples periods of zero demand (Kourentzes, 2013). The most known, the Croston's method (Croston, 1972) is present in commercial software, for example, SAP and Oracle RDF. In the CR method, the estimates are obtained as follows:

$$z_t = \alpha y_t + (1 - \alpha)z_{t-1} \tag{6}$$

$$\hat{p}_t = \alpha p_t + (1 - \alpha)\hat{p}_{t-1} \tag{7}$$

$$\hat{p}_t = \alpha p_t + (1 - \alpha)\hat{p}_{t-1} \tag{8}$$

The method consists of forecasting separately a value of demand,  $y_t$  in Equation 6, and the time interval between demands,  $p_t$  in Equation 7, assuming that both variables are independent (Croston, 1972). Finally, Equation 8 provides a rate of expected demand (forecast demand) by a period.

The AR models are a flexible class of models and can handle a wide range of time series patterns; but, in general, they are applied on stationary time series (Hyndman & Athanasopoulos, 2018). Unlike the traditional regression models, the independent variable is estimated by considering its past values (the autoregression term is used for this reason). The AR method of order  $p$ , also referred as  $AR(p)$ , was selected in this work and it can be written as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \tag{9}$$

In the Equation 9,  $c$  is a constant,  $\varepsilon_t$  is the white noise of the time series, and  $\phi_1, \dots, \phi_p$  are weight parameters. Note that the estimates (forecasts) are produced by a linear combination of lagged values of  $y$ .

## 2.2. Evaluation metrics for forecasting methods

To measure the accuracy of forecasting methods, several evaluation metrics can be found in literature. Mean Absolute Percentage Error (MAPE) is a popular evaluation metric, being suitable to evaluate different time series, because it is independent of the data scale and has easy interpretation. However, it can produce infinite or undefined errors if zero values (or approximately zero values) occur on the data, due division for the real value of demand. Since zero values are common in intermittent demand time series, MAPE is not suitable for this type of series as pointed in (Teunter & Duncan, 2009; Hyndman & Koehler, 2006; Makridakis et al., 2018).

Root Mean Square Error (RMSE) is a typical error metric widely employed in forecasting methods and machine learning methods. It does not suffer from the problem mentioned above. However, RMSE is more sensitive to outlier values. RMSE can be obtained as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (e_t)^2} \tag{10}$$

In the Equation 10,  $N$  is the number of observations, and  $e_t$  is a forecast error on time  $t$  and obtained as  $y_t - \hat{y}_t$ .

A survey of evaluation metrics for forecasting methods is proposed by Hyndman & Koehler (2006). Mean Absolute Scaled Error (MASE) can overcome the drawbacks of other evaluation metrics. This is because, MASE is independent of the data scale, less sensitive to outlier values and only produces undefined errors when all the NF forecast errors (on the denominator) are equals. Using a scaled error,  $q_t$ , as in Equation 11:

$$q_t = \frac{e_t}{\frac{1}{N-1} \sum_{i=2}^N |y_i - y_{i-1}|} \tag{11}$$

MASE is calculated by  $\text{mean}(|q_t|)$ , where  $N$  is the time series size on the training interval; and  $y_i$  and  $y_{i-1}$  are the real demand values on time  $i$  and  $i-1$ , respectively.

### 2.3. Combination of forecasting methods

The combination of forecasting methods has become an important strategy in many forecasting works and has used as a benchmark in many applications (Makridakis et al., 2018). For example, Wang & Petropoulos (2016) propose and evaluate the use a strategy of two models combination, namely judgmental adjustment and statistical output. The strategy is compared to a number of other demand forecast approaches. In most cases, to create a set of forecasting models and to perform a simple average of the methods' outputs usually obtain better accuracy than to use a single forecasting method.

This paper proposes the use of two combination strategies for aggregating forecasting methods: *simple mean* and *weighted mean*. In the first strategy, the final forecast is obtained by averaging the methods' outputs; and in the second strategy, the final forecast is calculated by taking a weighted sum of the methods' outputs, where the weight of each method is determined using the method's error on a validation interval. This work uses the terms  $CF_m$  and  $CF_w$  for the combination of forecasting methods with simple mean strategy and for the combination of forecasting methods with weighted mean strategy, respectively.

Assuming  $m$  as the number of forecasting methods,  $o_j$  as the output (forecast) of the model  $j$  for any time instant,  $w_j$  as the weight of the model  $j$ , the combination output is given by the Equation 12:

$$F(\mathbf{w}, \mathbf{o}) = \sum_{j=1}^m w_j o_j \quad (12)$$

For the weighted mean strategy ( $CF_w$ ), the RMSE and MASE metrics on a validation interval are used to compute the methods' weights. Consider  $error_j$  as the error value (i.e. RMSE value or MASE value) of a method  $j$  on a validation interval, its weight  $w_j$  can be computed by Equation 13 (Soares et al., 2012):

$$w_j = \frac{\text{adjusted error}_j}{\sum_{k=1}^m \text{adjusted error}_k} \quad (13)$$

where the adjusted error <sub>$j$</sub>  is computed as:

$$\text{adjusted error}_j = 1 - \text{average error}_j \quad (14)$$

and in Equation 14, the average error <sub>$j$</sub>  is:

$$\text{average error}_j = \frac{\text{error}_j}{\sum_{k=1}^m \text{error}_k} \quad (15)$$

Therefore, the main idea of the  $CF_w$  strategy is to assign a weight for each forecasting method according to its performance on a validation interval. For the  $CF_m$  strategy, the methods have the same contribution in the system, so that their weights are equal and set to  $w_j = 1/m$  (for  $j = 1, \dots, m$ ).

In this paper, the forecasting methods for designing the combination systems (i.e.  $CF_m$  and  $CF_w$ ) are SES, HOLT, NF, AR and CR. Therefore, the number of forecasting models is 5, so that  $m=5$  for Equation 12, Equation 13 and Equation 15.

### 2.4. Selection strategies for forecasting methods

A selection strategy aims to choose the best forecasting method (from a set of forecasting methods) using some criteria (for example, the accuracy on a validation interval). Different criteria can be found in literature. For example, based on the time series characteristics, Syntetos et al. (2005) proposed a selection scheme using the average inter-demand interval and the squared coefficient of variation of demand size. This scheme uses cut-off values to select between CR and an approach called Syntetos and Boylan approximation. This framework was proposed for intermittent demand time series.

On the other hand, Moon et al. (2013) propose other type of selection strategy. In this case, a squared coefficient of variation, a correlation and an equipment group (an external variable) were employed as inputs to a logistic regression model, where the main purpose it to predict which forecasting method has superior performance. The proposed model was evaluated using areal data set containing spare part demands from the Korean Navy.

Other approach is to select a forecasting method based on the performance. This approach consists of selecting the best forecasting method on the previous periods (data samples) assuming that the selected forecasting method will be more suitable for the next periods (Wang & Petropoulos, 2016). In most cases, the time series is divided in three time intervals, which are employed as training (used for building the forecasting method), validation (used for selecting the best forecasting method) and testing (used for evaluating the best forecasting method on a future time interval), respectively.

The performance of the selection strategies can be evaluated in terms of the accuracy of the selected forecasting method. For example, Wang & Petropoulos (2016) compared the performance of five different forecasting strategies: a forecasting method using a statistical model, a forecasting method adjusted by an expert, a combination of two forecasting methods, selections of a forecasting method based on the accuracy or on the variance. The performance of the strategies is evaluated in terms of inventory system metrics and forecasting method accuracy.

On the other hand, Fildes & Petropoulos (2015) propose four rules to select a forecasting model, and then they are compared to simple combination and aggregate selection (selection of the best forecasting method for all the time series). The strategies were analyzed with a subset of the M3-Competition data set, a popular data set for time series forecasts. The rules incorporate different selection criteria: “[...] best in-sample fit, best validation performance for one-step-ahead forecast, best validation performance on a pre-defined forecast horizon  $h$  and best validation performance for all forecast horizons” (Fildes & Petropoulos, 2015, p. 1694). Other aspects are considered in this work, such as, the size of the pool of forecasting methods and the accuracy of the individual selection. According to the authors, aggregate selection can be preferred if the data contain more similar sub-populations, but the individual selection can be necessary when the considered methods in the pool are low correlated.

The effect of the selection strategy for forecasting methods by experts was analyzed by Petropoulos et al. (2018) using data from the M3-Competition. The performed experiment differs from the others works due to the inclusion of a human judgment on the selection of forecasts. Also, a combination of forecasting methods is employed for performance comparisons. The results suggest that the inclusion of a human judgment on forecasting systems can be a useful practice (Petropoulos et al., 2018). Hyndman & Khandakar (2008) proposed two automatic selection procedures for forecasting methods. The algorithms can select a forecasting method based on the Akaike’s Information Criterion for Exponential Smoothing models or based on a Step-wise procedure for an ARIMA model.

This paper proposes and analyzes the use of a selection strategy for choosing the best forecasting model, from a pool of forecasting models, based on the performance on a validation interval. That is, in the first step, a pool of seven forecasting methods (namely, SES, HOLT, NF, AR, CR,  $CF_m$  and  $CF_w$ ) are designed using a training interval. After, the forecasting methods are evaluated using a validation interval. And then, the model with the lowest error (MASE or RMSE) is selected as the final forecasting model. Finally, the selected model is designed using the training and the validation intervals, and is evaluated on the testing interval to analyze its performance on future samples. In the next sections, this proposed approach, with automatic selection of a forecasting method, is termed as “AUTO”. Figure 1 presents an overview of the described approach.

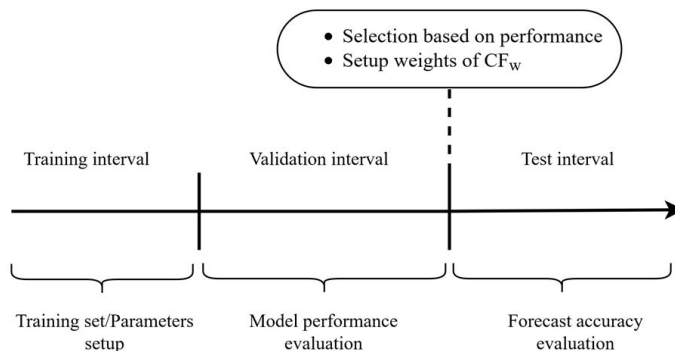


Figure 1. An overview of the time intervals in this paper.

### 3. Experimental design and results

In this section, the proposed forecasting methods ( $CF_m$ ,  $CF_w$  and AUTO) are evaluated using two real-world data sets: a spare part data set from an industry of elevators and the M3-Competition data set. The proposed approaches are compared to SES, HOLT, NF, AR and CR. The experiments were performed using the Python programming language, running on a PC equipped with an Intel® Core™ i5-7200U 2.50GHz processor of 2 cores and 8GB of RAM.



The implementations of the forecasting methods (except the CR method) can be found at a Python library (Statsmodels, 2020) called *Econometric and Statistical Modeling with Python* published by Seabold & Perktold (2010). On the other hand, the implementation of the CR method was developed using tools of the mentioned library and other Python libraries.

### 3.1. Data set description

The M3-Competition data set is a popular and public data set (M3-Competition, 2020) containing a large number of time series (Makridakis & Hibon, 2000). It consists of 3,003 time series, where 1,428 times series have monthly demand data of various types of applications, including industry, demographic, finance, among others. This data set was selected due to the large amount of observations (samples) on the time series, providing enough information to build forecasting methods.

The other data set is a private data set provided by an elevator industry organization in Brazil. It is a time series data with monthly demands of a spare part throughout the year 2019 in 54 geographic locations (cities) in Brazil. Therefore, the total of time series is 54, one for each geographic location. A particular characteristic of this data set is the high presence of demands with zero values. Therefore, most times series are intermittent.

A test to verify the intermittent characteristics of both data sets was performed, using the framework proposed Syntetos et al. (2005). The test consists of computing the values for the average inter-demand interval and the coefficient of variation. Then, it employs these values to classify a time series as intermittent, using cut-off values proposed by the authors. Using this test, it was verified that the M3-Competition data set does not have intermittent characteristics, whereas the spare part dataset has intermittent characteristics. Table 3 describes the main specifications of both data sets.

Table 3. Specification of the Data Sets Used in the Experiments.

Data set	Number of time series	Mean number of observations	Access type	Intermittent characteristics?
M3-Competition	1428	117	Public	No
Spare part	54	29	Private	Yes

Figure 2, 3 and 4 show selected time series from the M3-Competition data set, where Figure 2 shows a time series of size 30 ( $T=30$ ), Figure 3 presents a time series of size 60 ( $T=60$ ), and Figure 4 displays a time series of size 126 ( $T=126$ ). Times series from the spare part data set are omitted to preserve the privacy of this data set.

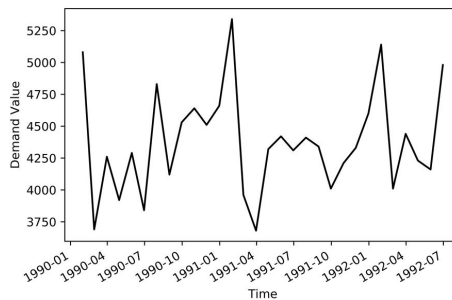


Figure 2. A time series of size 30 ( $T=30$ ) from the M3-Competition data set.



Figure 3. A time series of size 60 ( $T=60$ ) from the M3-Competition data set.



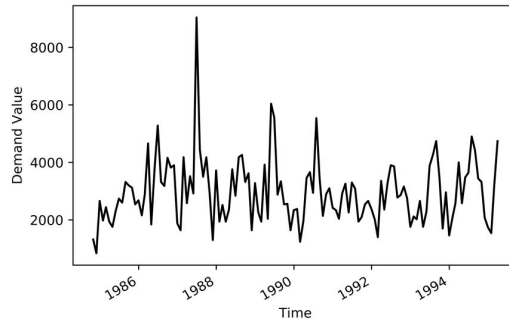


Figure 4. A time series of size 126 ( $T=126$ ) from the M3-Competition data set.

### 3.2. Approach description and setup

To train and evaluate the forecasting models, the following time series division was adopted. Each time series (of size  $T$ ) was divided into three time intervals: training interval (60%), validation interval (20%) and testing interval (20%). This division allows evaluating a time series according to its size, creating a more realistic scenario.

The first interval, with the training data, contains data from time 1 to  $T_1$ ; and it is used to fit and setup the forecasting models. The second interval, with the validation data, has data from time  $T_1+1$  to  $T_2$ ; and it is used to evaluate the performance of a forecasting method based using the predictions for this interval. The third interval, with the testing data, contains information from time  $T_2+1$  to  $T$  and will be used with a twofold purpose. That is, it will be used to evaluate the performance of the selection and combination strategies, and of the forecasting methods in terms of RMSE and MASE.

As described previously, to select the best forecasting method, the RMSE and MASE metrics on the validation data are used. Some authors name this approach as “past forecast performance” (Wang & Petropoulos, 2016; Fildes & Petropoulos, 2015). It means to produce a  $h$ -step-ahead forecast with  $h$  varying from 1 to  $T_2-T_1$ , which computes the accuracy and selects the method with the best value (lowest value) of RMSE or MASE, according to the configuration of an experiment.

The parameters and setup of the forecasting methods are the following:

- *Naive Forecast method* (NF). Implemented considering Equation 1. The NF method does not any require parameter setup; and it can be also implemented using the SES method (described below) by setting  $\alpha = 1$ ;
- *Simple Exponential Smoothing method* (SES). The selected smoothing parameter  $\alpha$  is the one that maximizes the log-likelihood. The first “in-sample” fitted value (i.e.  $\hat{y}_1$ ) is initialized using a grid search method;
- *Holt’s linear trend method* (HOLT). The smoothing parameter for level  $\alpha$  and the smoothing parameter for trend  $\beta$  are chosen by maximizing the log-likelihood. And, as in the SES method, a grid search method is used for initializing the level and the trend values;
- *Autoregressive method* (AR). The constant term  $c$ , the model order  $p$  and the coefficients  $\phi_1, \dots, \phi_p$  are estimated using unconditional maximum likelihood approach;
- *Croston Method* (CR). The smoothing parameter  $\alpha$  was set to 0.15, as suggested Teunter & Duncan (2009). The initialization values for the interval  $p$  are is first interval between demands; and the level  $z$  is the first non-zero value.

### 3.3. Evaluation methodology

To evaluate the performance of the forecasting methods, data from time 1 to  $T_2$  (training and validation intervals) are employed to train the methods and data from for the time  $T_2+1$  to  $T$  are used to evaluate the methods (testing interval). The  $h$ -step-ahead forecast for the testing interval will be produced by varying  $h$  from 1 to  $T-T_2$ . The RMSE and MASE errors of the forecasting methods on the testing interval are computed. This configuration allows to compare the performance of the selection procedure to the others forecasting methods.

Below, the results of the forecasting methods on the testing interval, averaged over all the time series are reported.

#### 4. Results and discussion

In this subsection, the results of the forecasting methods are reported. Table 4 presents the results of the M3-Competition data set using time series with 30, 60 and all the observations (i.e.  $T = 30$ ,  $T = 60$  and *Full*) using the RMSE and MASE metrics to compute the accuracy.

Table 4. Results of the M3-Competition Data Set with  $T = 30$ ,  $T = 60$  and *Full* (all the observations).

Forecasting Method	Average of RMSE on different values of $T$			Average of MASE on different values of $T$		
	$T = 30$	$T = 60$	<i>Full</i>	$T = 30$	$T = 60$	<i>Full</i>
AR	1149.536	883.593	871.413	2.585	2.346	2.510
AUTO	815.663	777.513	861.401	1.701	1.976	2.652
CF <sub>m</sub>	783.281	785.916	881.638	1.665	1.977	2.625
CF <sub>w</sub>	<b>776.355</b>	<b>773.046</b>	868.119	<b>1.635</b>	<b>1.931</b>	2.583
CR	795.502	805.393	927.457	2.275	2.651	3.473
HOLT	1036.631	1085.089	1167.305	1.906	2.325	2.876
NF	880.760	879.720	997.645	1.748	2.181	3.009
SES	792.204	804.115	916.457	1.647	2.102	2.931

By varying the time series sizes, it can be analyzed the effect of the  $h$  and the amount of used observations to produce the forecasts for a horizon  $h$ . Table 5 shows the results of the spare part data set. Each error value, for RMSE and MASE, is calculated by averaging the error values of all the time series. In all tables, the best performing method is highlighted in bold.

Table 5. Results of the Spare parts Data Set.

Forecasting Method	Average of RMSE	Average of MASE
AR	165.3232	1.397
AUTO	101.109	0.891
CF <sub>m</sub>	101.047	0.882
CF <sub>w</sub>	100.224	0.876
CR	95.940	<b>0.8443</b>
HOLT	128.434	1.199
NF	115.778	1.016
SES	100.429	0.870

Considering the time series with 30 observations (Table 4), the results reveal that the CF<sub>w</sub> method has the lowest error for MASE and RMSE. Therefore, the CF<sub>w</sub> method has good performance when compared to the other forecasting methods for small time series size. Moreover, the combination strategy using equal weights (CF<sub>m</sub>) has the second lowest error (considering RMSE). This shows that a combination of forecasting methods can outperform other approaches.

Considering the results shown in Table 4 ( $T=60$ ), the CF<sub>w</sub> method has good performance and is followed by the AUTO selection strategy, considering both metric errors. This result indicates that the CF<sub>w</sub> has good accuracy as the size of the time series increases. The CF<sub>m</sub> remains at the top of the three performing forecasting methods.

Table 4, *Full* column, shows the result using all the observations of the time series. In this case, the results present some differences. Regarding the AUTO method, its accuracy improved as the size of the time series increase, performing better than all the other forecasting methods, considering the RMSE metric. It suggests that the AUTO strategy is more sensitive to the size of the time series. The CF<sub>w</sub> method is also the best performing methods, achieving the second best accuracy in both metrics.

Moreover, AUTO and CF<sub>m</sub> have similar performance when considering the RMSE metric, but AUTO has worse performance when using the MASE metric. In general, all the methods increase the RMSE and MASE values as the time series size ( $T$ ) increases, since the forecasting horizon ( $h$ ) also increases (number of testing observations).

For example, the RMSE values of the  $CF_w$  method are 776.355, 773.046 and 868.119 for  $T=30$ ,  $T=60$  and Full, respectively. This occurs because when  $h$  is larger, more uncertainty is associated to the time horizon  $h$  [9]. In particular, the AR method may be less sensitive to this effect, since the RMSE values are 883.593 and 871.413 for  $T=60$  and full time series data, respectively.

The performance results of the spare part data set (Table 5) are similar to the results of the M3-Competition data set. The CR method, which is an approach for intermittent series, performs better than the other forecasting methods. In this case, CR has achieved 95.940 and 0.844 values for RMSE and MASE, respectively. Considering RMSE, the second best performing method is  $CF_w$ , which achieved good performance on this data set, but, in this case, it does not outperform CR. When comparing the performances of AUTO and  $CF_m$ , AUTO has 101.109 for RMSE and  $CF_m$  has 101.047 for RMSE; and, for MASE, AUTO has 0.891 for MASE and  $CF_m$  has 0.882 for MASE. Thus, the AUTO strategy has worse performance than  $CF_m$  with data from the spare parts data set.

Additionally, a notable performance of SES was obtained in both data sets, performing better than AUTO and  $CF_m$  in some cases. This confirms the popularity of SES in literature and among the providers of commercial software.

In general, the selection and the combination strategies have good generalization performance on both data sets, so that they can be efficiently applied to other data sets.

It should be pointed that the AUTO has the performance similar to  $CF_m$ , so that an additional test was performed to analyze their accuracy. Table 6 shows the percentage of time series in which AUTO has better accuracy (lowest error) than  $CF_m$  in the testing interval. For example, considering the M3-Competition data set (with "Full" times series size) and the RMSE metric, in 50.14% of the times series (from a total of 1,428 times series), AUTO outperforms  $CF_m$ . The results confirm that both forecasting methods have similar performance, considering data from spare parts data set and M3-Competition.

Table 6. Percentage of Better Accuracy of AUTO than  $CF_m$ .

Data set	Value (%)	Metric
M3-Competition	48.73	MASE
	50.14	RMSE
Spare part	42.59	MASE
	59.25	RMSE

## 5. Conclusion

This work proposes a forecasting application strategy considering two procedures, the combination of state-of-the-art forecasting methods and the selection of forecasting methods based on the models' accuracy. Two combination strategies are proposed: simple mean and weighted mean based on the methods' accuracy. This paper evaluates the model performance by using the MASE and RMSE metrics in order to measure the accuracy of the forecasting strategies under different scenarios, avoiding problems reported by previous works (Hyndman & Koehler, 2006).

To simulate different and more realistic scenarios, this work used two data sets with different characteristics, a public dataset of the M-Competition and a private data set of spare part demand from an elevator industry. This last data set presents a particular characteristic of time series called intermittency. The tested data sets allow assessing the generalization of the proposed strategies in other data sets.

The combination of forecasting methods demonstrates to be valuable if a weighting scheme based on the performance is employed. Although, the combination using simple mean outperforms other forecasting methods (such as, SES). The combination strategy is easy to understand and implement, and can be used in future works of forecasting methods. Moreover, the experiment results indicate that the automatic selection strategy based on the performance on a validation interval (AUTO) may not be good criteria for selecting forecasting models, since  $CF_w$  outperforms AUTO.

In general, the results suggest that combination strategies have potential application in demand forecasting problems, outperform other state-of-the-art models in trend and stationary series, and have comparable accuracy to other models in intermittent series. Therefore, they can be used to improve production planning activities in different applications and scenarios. Therefore, future works should be devoted to test other selection criteria. For example, a selection strategy based on the inventory performance, as proposed by Wang & Petropoulos (2016). Moreover, future works can also consider using a rolling (dynamic window) forecast design (Fildes & Petropoulos, 2015; Wang & Petropoulos, 2016). The inclusion of other forecasting methods (such as,

multivariate methods and machine learning methods) in the pool of selection and combination models can be also considered as a future work.

## Acknowledgements

This work is being partially supported by European Union's Horizon 2020 research and innovation programme and from Brazilian Ministry of Science, Technology and Innovation through Rede Nacional de Pesquisa under the Grant Agreement 777096, project "FASTEN - Flexible and Autonomous Manufacturing Systems for Custom-Designed Products".

## References

- Adya, M., Collopy, F., Armstrong, J. S., & Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting*, *17*(2), 143–157. [http://dx.doi.org/10.1016/S0169-2070\(01\)00079-6](http://dx.doi.org/10.1016/S0169-2070(01)00079-6).
- Armstrong, J. S. (Ed.) (2001). *Principles of forecasting* (Vol. 30). USA: Springer. <http://dx.doi.org/10.1007/978-0-306-47630-3>.
- 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. <http://dx.doi.org/10.1016/j.ijpe.2018.01.026>.
- Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. <http://dx.doi.org/10.1016/j.ijpe.2016.03.017>.
- Choi, J. Y., & Lee, B. (2018). Combining LSTM Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting. *Mathematical Problems in Engineering*, *2018*, 1–8. <http://dx.doi.org/10.1155/2018/2470171>.
- Collopy, F., & Armstrong, J. S. (1992). Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, *38*(10), 1394–1414. <http://dx.doi.org/10.1287/mnsc.38.10.1394>.
- Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly (1970-1977)*, *23*(3), 289–303. <https://doi.org/10.2307/3007885>.
- Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701. <http://dx.doi.org/10.1016/j.jbusres.2015.03.028>.
- Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: implementation patterns in manufacturing companies. *International Journal of Production Economics*, *210*, 15–26. <http://dx.doi.org/10.1016/j.ijpe.2019.01.004>.
- Franses, P. H., & Legerstee, R. (2011). Combining SKU-level sales forecasts from models and experts. *Expert Systems with Applications*, *38*(3), 2365–2370. <http://dx.doi.org/10.1016/j.eswa.2010.08.024>.
- Guo, F., Diao, J., Zhao, Q., Wang, D., & Sun, Q. (2017). A double-level combination approach for demand forecasting of repairable airplane spare parts based on turnover data. *Computers & Industrial Engineering*, *110*, 92–108. <http://dx.doi.org/10.1016/j.cie.2017.05.002>.
- Heinecke, G., Syntetos, A. A., & Wang, W. (2013). Forecasting-based SKU classification. *International Journal of Production Economics*, *143*(2), 455–462. <http://dx.doi.org/10.1016/j.ijpe.2011.11.020>.
- Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. <http://dx.doi.org/10.1016/j.ijforecast.2003.09.015>.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Melbourne, Australia: OTexts. Retrieved in 2020, January 31, from <https://otexts.com/fpp2/>
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22. <http://dx.doi.org/10.18637/jss.v027.i03>.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. <http://dx.doi.org/10.1016/j.ijforecast.2006.03.001>.
- Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. *International Journal of Production Economics*, *143*(1), 198–206. <http://dx.doi.org/10.1016/j.ijpe.2013.01.009>.
- Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226–235. <http://dx.doi.org/10.1016/j.ijpe.2018.05.019>.
- M3-Competition. (2020). *M3-Competition - International Institute of Forecasters*. Retrieved in 2020, January 31, from <https://forecasters.org/resources/time-series-data/m3-competition/>.
- Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. [http://dx.doi.org/10.1016/S0169-2070\(00\)00057-1](http://dx.doi.org/10.1016/S0169-2070(00)00057-1).
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808. <http://dx.doi.org/10.1016/j.ijforecast.2018.06.001>.
- Moon, S., Simpson, A., & Hicks, C. (2013). The development of a classification model for predicting the performance of forecasting methods for naval spare parts demand. *International Journal of Production Economics*, *143*(2), 449–454. <http://dx.doi.org/10.1016/j.ijpe.2012.02.016>.
- Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*(1), 34–46. <http://dx.doi.org/10.1016/j.jom.2018.05.005>.
- Qi, M., & Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, *132*(3), 666–680. [http://dx.doi.org/10.1016/S0377-2217\(00\)00171-5](http://dx.doi.org/10.1016/S0377-2217(00)00171-5).
- Rego, J. R., & Mesquita, M. A. (2015). Demand forecasting and inventory control: a simulation study on automotive spare parts. *International Journal of Production Economics*, *161*, 1–16. <http://dx.doi.org/10.1016/j.ijpe.2014.11.009>.

- Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* (pp. 92-96). USA: SciPy.org. <http://dx.doi.org/10.25080/Majora-92bf1922-011>.
- Soares, S., Antunes, C., & Araújo, R. (2012). A genetic algorithm for designing neural network ensembles. In *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference - GECCO '12* (pp. 681-688). Canadá: ACM. <http://dx.doi.org/10.1145/2330163.2330259>.
- Statsmodels. (2020). *StatsModels: Statistics in Python - statsmodels v0.10.1 documentation*. Retrieved in 2020, January 31, from <https://www.statsmodels.org/v0.10.1/#>.
- Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *The Journal of the Operational Research Society*, *56*(5), 495-503. <http://dx.doi.org/10.1057/palgrave.jors.2601841>.
- Teunter, R. H., & Duncan, L. (2009). Forecasting intermittent demand: a comparative study. *The Journal of the Operational Research Society*, *60*(3), 321-329. <http://dx.doi.org/10.1057/palgrave.jors.2602569>.
- Wang, X., & Petropoulos, F. (2016). To select or to combine? The inventory performance of model and expert forecasts. *International Journal of Production Research*, *54*(17), 5271-5282. <http://dx.doi.org/10.1080/00207543.2016.1167983>.
- Yu, Y., Choi, T.-M., & Hui, C.-L. (2011). An intelligent fast sales forecasting model for fashion products. *Expert Systems with Applications*, *38*(6), 7373-7379. <http://dx.doi.org/10.1016/j.eswa.2010.12.089>.