UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

DEPARTEMENT D'INFORMATIQUE
DEPARTEMENT FÜR INFORMATIK
Bd de Pérolles 90
CH-1700 Fribourg

www.unifr.ch/inf

WORKING PAPER

How to improve accessories sales forecasting of a medium-sized Swiss enterprise? A comparison between statistical methods and machine learning algorithms

A. Ramosaj, N. Ramosaj, M. Widmer

Internal working paper no 24-01

May 2024

How to improve accessories sales forecasting of a medium-sized Swiss enterprise? A comparison between statistical methods and machine learning algorithms

Agneta Ramosaj[a,*], Nicolas Ramosaj[b], Marino Widmer[a]

[a] *Department of Informatics - Decision Support and Operations Research, University of Fribourg, Switzerland*
[b] *School of Engineering and Architecture of Fribourg, HES-SO University of Applied Sciences and Arts, Western Switzerland*
*Corresponding author: agneta.ramosaj@unifr.ch

ARTICLE INFO

Keywords:
Demand forecast
Key account manager (KAM)
Seasonal autoregressive integrated moving average (SARIMA)
Machine learning (ML)
K-nearest neighbors (k-NN)
Least absolute shrinkage and selection operator (LASSO)
Linear regression
Random forest (RF)
Root mean square error (RMSE)
Mean absolute error (MAE)

ABSTRACT

Forecast accuracy is a crucial topic for industrial companies, and its impacts are particularly important for the finance and production departments. The company can incur high costs if forecasts are not accurate, for example, due to stock-outs or excess inventory.
Therefore, the purpose of this study was to optimize accessories forecasting for a medium-sized Swiss enterprise. To do so, different forecasting techniques were tested, and statistical methods and machine learning (ML) algorithms were compared. The results were adjusted according to key account managers' (KAM) expertise.
This paper presents a comparison between exponential smoothing, seasonal autoregressive integrated moving average (SARIMA), SARIMAX (SARIMA with exogenous variables) and ML algorithms, such as k-nearest neighbors (k-NN), least absolute shrinkage and selection operator (LASSO) regression, linear regression, and even random forest (RF).
To compare these different methods, two measures of statistical dispersion are computed: mean absolute error (MAE) and root mean squared error (RMSE). The results are standardized to enable a better comparison. For our dataset, SARIMAX (with the KAMs' expertise as an exogenous variable) appears to give better results than all the ML algorithms tested.

1. Introduction

A previous article focused on improving product forecast accuracy for a Swiss small and medium enterprise (Ramosaj A., 2022): it identifies certain good methods to improve product forecasting accuracy. The following article discusses how forecasting accuracy can also be improved for accessories and spare parts. The difference between final products and accessories/spare parts is that products refer to household products while accessories are necessary for household products to be used and include, for example, spare parts, esthetic elements, or even maintenance products. Bad forecasting for spare parts can negatively impact decisions on supply, additional storage and maintenance costs, destruction, etc.

There are noticeable differences between forecasting methods for final products and spare parts (Morris M., 2013). Various studies have been carried out to improve spare parts forecasting. Most try to reduce inventory costs (Romeijnders W., 2012), while others attempt to make links between poor accuracy and inventory obsolescence (Teunter R.H., 2011). Different forecasting approaches have been applied to improve the accuracy of spare parts orders. An interesting review of the field by (Pinçea C., 2021), splits spare parts forecasts into three categories: (1) time-series methods; (2) a combination of contextual approaches and statistical forecasting techniques; and (3) comparison of traditional and alternative demand forecasting methods.

The consumption of spare parts varies greatly and is therefore challenging to forecast. Some research has focused on investigating the gap between research and practice in spare parts management through different case studies (Bacchetti A., 2012), while other studies have focused on statistical analysis (Hemeimat R., 2016). An interesting paper by (Arvan M., 2019) introduced human adjustments to quantitative forecasts, while another study made links between judgmental adjustments and statistical methods (Van den Broeke M., 2019). (Hu Q., 2018) focused on spare parts classification by defining criteria such as the repairability, lead time, and even obsolescence of spare parts. Most studies have used historical data to determine how best to improve spare parts and applied models such as ARMA, auto-regressive integrated moving average (ARIMA) (Sheng F., 2020), autoregressive integrated moving average with exogenous values (ARIMAX), and seasonal autoregressive integrated moving average with exogenous values (SARIMAX) (Arunraj N.S., 2016). To go further, others have compared statistical and machine learning (ML) methods, such as k-nearest neighbors (k-NN) or even random forest (RF) (Spiliotis E., 2020). Most studies have used accuracy to measure the performance of forecasting, but others have shown that inventory performance yields more realistic benchmarks (Teunter R.H., 2017).

The aim of our study was to improve the forecasting of spare parts for a medium-sized Swiss enterprise by using time-series models as well as ML algorithms. Following (Xie M., 2013), it was decided to add exogenous variables to the seasonal auto-regressive integrated moving average (SARIMA). The objective was to check the impacts of different exogenous values on the SARIMA model and compare these results with ML algorithms. We chose SARIMA, SARIMAX (SARIMA with exogenous variables), and some ML algorithms, such as LASSO, k-NN, RF, and linear regression, as exponential smoothing models.

To outline the performance of our results, we decided to combine traditional forecasting measures, such as mean absolute error (MAE) and root mean square error (RMSE) but also residual stock (RS), to check the inventory impacts, as suggested in (Teunter R.H., 2017).

This article is organized as follows. Section 2 presents the state of the art of some forecasting models, after which Section 3 outlines the methodology used to solve the present research problem. Section 4 discusses the results and comparisons, and the last section concludes the paper.

2. Existing forecasting models

2.1 Exponential Smoothing

(Stadtler H., 2002) defined exponential smoothing as the most frequently used forecasting method. The equation is the following:

$$F_t = F_{t-1} + \alpha\,(A_{t-1} - F_{t-1}) \tag{1}$$

where F_t is the forecast value for period t, F_{t-1} is the exponentially smoothed forecast made for the previous period, A_{t-1} is the actual demand in the previous period and alpha (α) is a smoothing constant that provides the weight of the committed error.

2.2 SARIMA

ARIMA is frequently used in spare parts forecasting (Jiafu R., 2009). The initial parameters used in ARIMA (p, d, q) are auto-regression [AR(p)], integrated [I(d)], and moving average [MA(q)].

The difference between ARIMA and SARIMA is that the second includes the seasonality effects (Hyndman R.J., 2018). The equations of ARIMA and SARIMA are explained in (Ramosaj A., 2022).

SARIMA can be expressed as follows:

$$\text{SARIMA}\ \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}}\ \underbrace{(P,D,Q)\text{s}}_{\text{Seasonal part of the model}}, \tag{2}$$

where s is the number of observations in which uppercase notation can be observed and defined as the seasonal parts (s) of the model, as depicted in Equation 3.

$$(1-\text{Ø}_1\text{B})\ (1-\underbrace{\phi_1\text{B}^{\text{s}}}_{\text{Seasonal AR(1)}})\underbrace{(1-\text{B})}_{\text{Non-seasonal difference}}\underbrace{(1-\text{B}^{\text{s}})}_{\text{Seasonal difference}}y_t=(1+\theta_1\text{B})\ (1+\underbrace{\Theta_1\text{B}^{\text{s}}}_{\text{Seasonal MA (1)}})\varepsilon_t. \tag{3}$$

Non-seasonal AR (1) — Non-seasonal MA (1)

P,D,Q and p,d,q are sets based on the best Akaike's Information Criterion (AIC). AIC is defined as: AIC = −2log(L) + 2K, where L denotes the likelihood and K the number of parameters estimated by the model. The AIC penalizes models with many parameters, and thus attempts to select the best model by favoring simpler models. Note that the AIC does not have much meaning by itself. It is only useful in comparison to the AIC value for another model fitted to the same dataset (Makridakis S., 2008).

2.3 SARIMAX

To go further and include the effect of demand influencing factors, SARIMAX is applied. This model is a combination of the SARIMA model and an X factor that represents an exogenous value (Arunraj N.S., 2015). This exogenous value, X, is an external factor that impacts the accuracy of the forecasts. SARIMA (Ramosaj A., 2022) could be expressed by (p,d,q) (P,D,Q)s. To move from SARIMA to SARIMAX, an additional external factor, X, is added and is modeled by multi-linear regression equation as follows:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \omega_t \quad (4)$$

where:
y_t is the time series dependent variables.
$x_{1,t}\ to\ x_{k,t}$ are the external variables.
$\beta_0\ to\ \beta_k$ are the correlation regression coefficients.
ω_t is a stochastic residual independent of the input series.

ω_t can be formulated as follows:

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_p(B^S)(1-B)^d(1-B^s)^D}\,\varepsilon_t \quad (5)$$

where:
$\theta_q(B)$ is the non-seasonal moving average (MA).
$\Theta_Q(B^s)$ is the seasonal moving average (MA).
$\phi_p(B)$ is the non-seasonal auto-regression (AR).
$\Phi_p(B^S)$ is the seasonal auto-regression (AR).
$(1-B)^d$ is the non-seasonal difference (I).
$(1-B^s)^D$ is the seasonal difference (I).
ε_t is the remaining error.

The general formula of SARIMAX can be obtained by substitution, namely, by putting equation 5 in equation 4 as follows (Arunraj N.S., 2016):

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_p(B^S)(1-B)^d(1-B^s)^D}\,\varepsilon_t \quad (6)$$

Each exogenous value is appropriate to each spare part forecasted. We decided to use the following exogenous variables in our study:

- Month sales: indication of the months; for example, January is 1 and August is 8.

- Mean sales: mean of all August months for the four years of data, for example:

$$\frac{\text{Sales. August 2018} + \text{Sales. August 2019} + \text{Sales. August 2020} + \text{Sales. August 2021}}{4}$$

- KAM: the given key account managers' forecasts.
- Mean KAM: mean of all August months predicted by the KAM for the four years of data, for example:

$$\frac{\text{KAM. August 2018} + \text{KAM. August 2019} + \text{KAM. August 2020} + \text{KAM. August 2021}}{4}$$

- Last year KAM: last year's KAM forecasts based on 12 months' rolling values.
- Regression KAM: linear regression on all the KAM predictions used to find the values of forecasts for the next year.

2.4 ML models and metrics

ML in forecasting uses algorithms to learn from past sales to predict future events. While statistical methods use linear processes to forecast, ML uses non-linear algorithms. ML methods have become popular with the increase in interest in artificial intelligence (AI) (Makridakis S., 2018). For the data set used in this study, the most common ML algorithms were used, such as least absolute shrinkage and selection operator (LASSO), k-NN, RF, and linear regression.

2.4.1 LASSO

LASSO regression measures the relationship between variables. The goal is to find a balance between the best accuracy and overfitting by adding a penalty term to the traditional linear regression models (Sethi J.K., 2021). The parameters will be close to zero if the influence on the prediction model is low. LASSO is helpful when many variables are used in a model but not all of them are relevant (Ranstam J., 2018).

The objective function is to minimize the following equation:

$$\mathrm{J(w)} = \frac{1}{2m}\left[\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{n}|w_j|\right] \quad (7)$$

where y_i is the observed values (test set), and $\hat{y}_i$ is the predicted values.

This equation is based on two parts:

First part:

$\frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$: the goal is to minimize the prediction error.

Second part:

$\lambda \sum_{j=1}^{n}|w_j|$: the goal is to minimize the value of the parameters, where λ is a constant and $|w_j|$ is the L_1-norm of the coefficient vector. λ is a regularization parameter which is a hyperparameter between 0 and 1. It refers to the degree of penalty that is assigned to each

parameter of the model.

2.4.2 k-NN

k-NN regression is a nonparametric method (which means it does not make any assumptions about the underlying data) that bases its prediction on feature similarities. This type of regression is used when data are labeled or noise-free. k-NN is a "lazy learner": it does not learn from the training set immediately.

k is a parameter that refers to the number of nearest neighbors to include in most of the processes. Choosing the right value of k is important for accuracy. k can be chosen by taking two steps:

- Sqrt (n), where n is the total number of data points.
- An odd value of k, to avoid confusion between the data classes.

To obtain the nearest neighbors, we computed the Euclidean distance between the given point and all the points of the training set. The formula is the following:

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (8)$$

where x_i and y_i are the two points for which the distance is calculated.

2.4.3 Random Forecast

RF is built by taking a collection of multiple random decision trees. As it uses multiple different decision trees, RF is less sensitive to the training data than other ML models. The sample of each tree is taken randomly by bootstrapping (Tugay R., 2020). Bootstrapping helps to create subsets of the original dataset with replacement. At every sequential process, each model tries to correct the errors of the previous model.
The following steps are taken to build the RF:

- As mentioned, the first step is to create subsets of the original data, as shown in Figure 1. Rows and columns are selected through replacement.
- An individual decision tree is created for each subset.
- The decision trees all give a different output.
- The majority voting is combined in the final output.

The more different decisions trees are used, the better the accuracy. A benefit of the RF algorithm is that it can help reduce overfitting and bias (Mei J., 2014). To set up the RF, certain parameters must be set, namely, node, size, number of trees, and number of features.

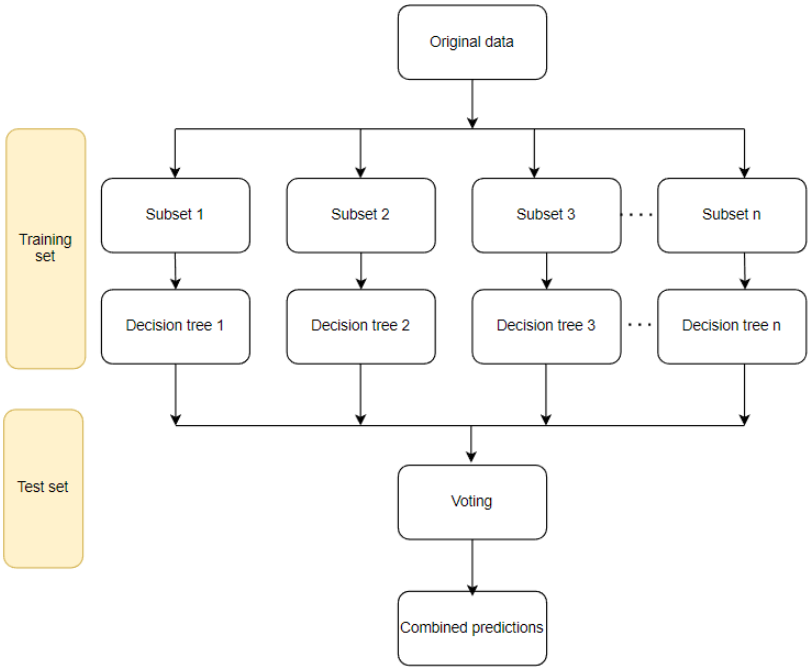

Figure 1: RF structure (designed by the authors)

To build the RF algorithm, the following steps must be followed (Noureen S., 2019):

a. For s = 1 to S = forest size

- Draw a bootstrap sample Z of size n from the training data, as shown in Figure 1.
- Grow an RF tree T_s to the bootstrapped data. To grow the tree, it is necessary to repeat the following steps at each node until the minimum node size n_{min} is reached.
    1. Select the number of randomly chosen factor r from the p variables.
    2. Pick the best p among the r.
    3. Split the node into two sub-nodes.

b. Output the ensemble of the trees $\{T_s\}_1^S$.

To make a prediction at a new point, x, we used the following formula for the RF algorithm based on regression:

$$\hat{f}_{rf}^{S}(x) = \frac{1}{S}\sum_{s=1}^{S} T_s(x) \tag{9}$$

RFs aim to improve the variance reduction of bagging by reducing the correlation between the trees without overly increasing the variance. Bagging is a technique that helps to reduce the variance within a noisy dataset. It is possible to achieve it in the tree-growing process by a random selection of the inputs. Hence, when a tree is growing on a bootstrapped dataset, a check must be made before each split to determine the selected factor r ≤ p of the input variables.

$$\rho\sigma^2 + \frac{1-\rho}{S}\sigma^2 \tag{10}$$

where:
ρ is the positive pairwise correlation.
σ^2 is the variance.

2.4.4 Linear regression

Linear regression is an important algorithm that assumes a linear connection between the independent variable x_i and the dependent variable y_i. It uses this linear connection to assume a relationship between these two variables (James G., 2021). Linear regression aims to find the best value of a and b, which means the error between the predicted value and the independent variable should be as low as possible.

$$y_i = a + bx_i + e_i \tag{11}$$

where:

$y_i =$ predicted value
$a =$ regression constant
$b =$ regression coefficient
x_i = independent variable or observed variable (the variable we expect to be influencing y)
e_i = error
i = 1,…, n

2.4.5 Standardization

The goal of data standardization is to convert data into a common format to enable better processing and comparison (Berner R., 2019). Data standardization can enhance accuracy and help in selecting the most relevant forecasting model (Grannis S.J., 2019). Therefore, standardization is carried out as a data pre-processing step before data are used in a time-series model or an ML model. To standardize the data, the following points must be observed:

- Numerical data should be on a consistent scale.
- The categories should be consistently named.
- The data value should be in a consistent format.
- A uniform approach must be taken to deal with undefined or missing values.

Zero-normalization (z-norm) is the most common method of standardizing data. The formula is as follows:

$$\mathrm{x}' = \frac{(\mathrm{x}-\mu)}{\sigma} \tag{12}$$

where:
x' is the standardized value.
x is the observed value.
N_{train} is the number of sales values used to train the model.

$\mu = \frac{1}{N_{train}} \sum_{n=1}^{N_{train}} x_n$ is the average deviation for the given feature.

$\sigma = \sqrt{\frac{1}{N_{train}} \sum_{n=1}^{N_{train}} (x_n - \mu)^2}$ is the standard deviation for the given feature.

After standardization, all the features will have a mean of zero and a standard deviation of one and respectively the same scale (Milligan G.W., 1988).

2.4.6 Dataset preparation

To make a prediction with an ML model, it is necessary to process the time-series into a matrix of history. To do so, multiple retrospective observations must be defined, and a target value must be labeled, such as the next month after the history of observations is considered. Conversely, statistical models are based only on the time-series itself and do not require any processing. However, in both cases, the temporal aspect of the series must be preserved for the best predictions to be made. Figure 2 and Table 1 and Table 2 are illustrative examples. The chart presents the time-series as a temporal chart, and the data are organized by date from January 2018 to December 2022. We defined a window of observations as 12, and we built the matrix as shown in Table 1. We kept the observations from January 2018 to December 2018 and stored the value that the ML model needed for the training in column F(t+1), where F was used for the forecast, t for current time, and 1 for the next month. Then, we continued with the second row, where the window moves from 1, and reperformed the process until the whole time-series had been parsed. Then, the dataset was shuffled by rows, and the ML model was trained.

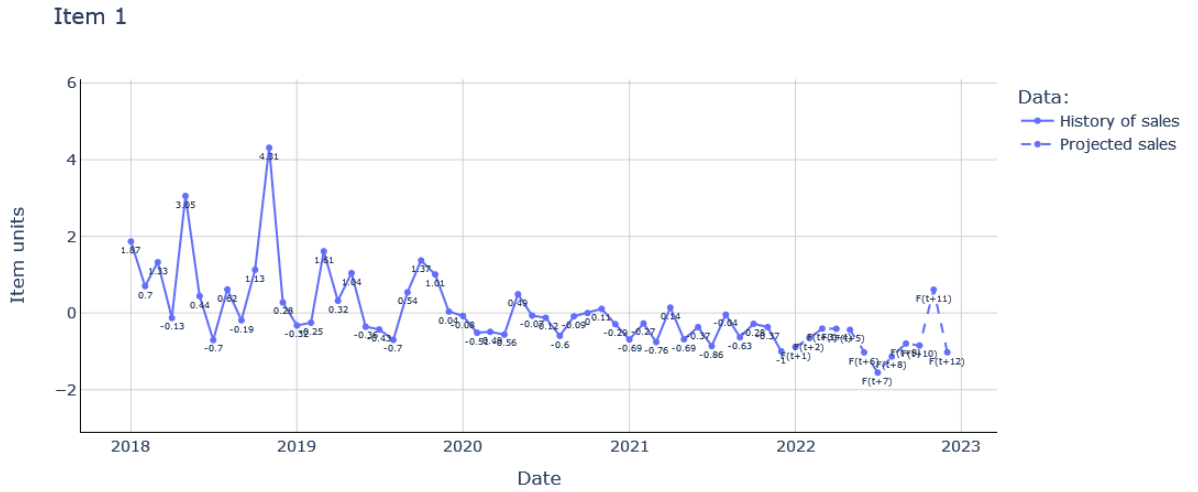

Figure 2: Normalized time-series for item 1 with displayed values. The last 12 months will be predicted and are represented as F(t+n), where F is used for forecast, t for current time, and n for the future.

	hist1	hist2	hist3	hist4	hist5	hist6	hist7	hist8	hist9	hist10	hist11	hist12	F(t+1)
1	1.87	0.7	1.33	-0.1	3.05	0.44	-0.7	0.62	-0.2	1.13	4.31	0.28	-0.32
2	0.7	1.33	-0.1	3.05	0.44	-0.7	0.62	-0.2	1.13	4.31	0.28	-0.32	-0.25
3	1.33	-0.1	3.05	0.44	-0.7	0.62	-0.2	1.13	4.31	0.28	-0.32	-0.25	1.61
4	-0.1	3.05	0.44	-0.7	0.62	-0.2	1.13	4.31	0.28	-0.32	-0.25	1.61	0.32
5	3.05	0.44	-0.7	0.62	-0.2	1.13	4.31	0.28	-0.3	-0.25	1.61	0.32	1.04

Table 1: Dataset created for the use of ML models based on Figure 2. A history of 12 observations was built with a targeted value of F(t+1), where F is used for forecast, t for current time, and 1 for the next month.

Once the ML model had been trained, it was ready to make predictions for the future. However, the ML model could only predict the next month, and the objective was to predict the next 12 months. Therefore, we resumed the predictions already made to predict the future. Table 2 shows the process of making predictions for the next 12 months. The predictions were based on a rolling window that considers a history as 12 months and predicts one month ahead. After 12 iterations of predictions, the forecasts for the next 12

months were returned and could be compared through metrics.

	hist1	hist2	hist3	hist4	hist5	hist6	hist7	hist8	hist9	hist10	hist11	hist12	F(t+1)
1	-0.69	-0.27	-0.76	0.14	-0.69	-0.37	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)
2	-0.27	-0.76	0.14	-0.69	-0.37	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)
3	-0.76	0.14	-0.69	-0.37	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)
4	0.14	-0.69	-0.37	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)
5	-0.69	-0.37	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)
6	-0.37	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)
7	-0.86	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)	F(t+7)
8	-0.04	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)	F(t+7)	F(t+8)
9	-0.63	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)	F(t+7)	F(t+8)	F(t+9)
10	-0.28	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)	F(t+7)	F(t+8)	F(t+9)	F(t+10)
11	-0.37	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)	F(t+7)	F(t+8)	F(t+9)	F(t+10)	F(t+11)
12	-1	F(t+1)	F(t+2)	F(t+3)	F(t+4)	F(t+5)	F(t+6)	F(t+7)	F(t+8)	F(t+9)	F(t+10)	F(t+11)	F(t+12)

Table 2: Predictions of the ML model based on a rolling window. Twelve iterations of predictions are necessary to return the next 12 months of forecasts as the columns F(t+1), where F is used for forecast, t for current time and 1 for the next month.

The performances were also benchmarked by using the KAM forecasts in the same way to increase the number of experiments for the training of the ML model (only used in Table 1).

2.4.7 Models' adaptation

To fit a statistical or ML model, the data must be adapted to preserve the integrity and consistency of the time-series. The data processing, model building, and estimation of the metrics were carried out using the Python libraries Scikit-learn and Statsmodels. The model fitting was still based on data from January 2018 to December 2021. The data from January 2022 to December 2022 were set aside for the evaluation of the forecasts.

Regarding the exponential smoothing model, the data were considered as a time-series format, the alpha parameter was auto-adjusted, and the series was directly given to the model.

In the SARIMA model, the data were preserved as a time-series format, and the parameters, such as p,d,q for the ARIMA model and P,D,Q, and s for the seasonality, were set based on the best AIC computed on the training set and a mapping of the different previously mentioned parameters. The model with the lowest AIC was used to make the forecasts. For SARIMAX, the method was the same as for SARIMA except that the exogenous values were added to the model during the training and forecasts.

The different ML models required the matrix with the history of the 12 values to forecast the next month (Table 1) as input. The ML models were trained by searching for the best parameters based on the training set itself, which consisted of the cross-validation of the training set to find the best parameters of the model. The method consisted of splitting the dataset into batches, also known as folds, and performing the training on the number of folds minus one and forecasting on the one remaining fold. The process was repeated on the same number of iterations as the number of folds. In our case, we considered five folds for all items and all ML models; hence, for every item, five iterations per parameter were

mapped through different values.

For instance, the LASSO model needed to be fitted through the alpha parameter only, and the alphas to map were 0.001, 0.01, 0.1, and 1. For every parameter's value, five training rounds were required due to the quantity of folds, resulting in 20 iterations for the four different values mapped (5 x 4 = 20). The complexity increased when the k-NN model was used, for which two parameters were mapped, namely, the number of neighbors and the methods to compute the distance. The number of neighbors consisted of three values (2, 3, 5), and the distance could be Manhattan or Euclidean (1 or 2). We kept the number of folds at five and, due to the mapping of the parameters, we had six couples of parameters to test by item, resulting in 30 iterations to select the best model (2 x 3 x 5 = 30).

Regarding the RF, the number of parameters increases again when considering the number of trees or estimators with five values (5, 10, 15, 20, 25) to map, the maximum depth of the tree with a mapping between five values (2, 4, 6, 8, 10), the minimum number of samples required to split a node to map at three values (2, 3, 5), and the bootstrap, which is a method to use the entire dataset to train all the trees at once, or the dataset is split proportionally to the number of trees so that every tree is trained with unique data. The bootstrap is mapped as two values (true or false). Based on the previous calculation, the RF needed 750 iterations to find the best parameters for the model (5 x 5 x 3 x 2 x 5 = 750).

Finally, the linear regression model does not need to search for the best parameters of the model because this model minimizes the residual sum of square without any restrictions or constraints. The ML models could be trained with an enlarged dataset thanks to the KAM forecasts, which were added to the training set.

Once the ML model's parameters were found, the final model could be trained based on the entire training set, and the forecasts could be made as illustrated in Table 2. Metrics were applied to measure the performances of the models.

2.4.8 Metrics

To measure the performance of our different models, we chose RMSE, MAE, and the RS as metrics. Using the RS enabled the inventory impacts to be checked, as suggested in (Teunter R.H., 2017).

2.4.8.1 RMSE

RMSE measures the magnitude of the error, and better models have a lower RMSE (Bakay M.S., 2021). The formula is the following:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (13)$$

In the equation, n is the number of data points or observations, y_i is the observed values (test set), and $\hat{y}_i$ is the predicted values.

2.4.8.2 MAE

The MAE is the sum of the absolute residual error. The formula is:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \quad (14)$$

2.4.8.3 RS

RS measures whether the forecasts are too optimistic (high stock level) or pessimistic (out of stock [OOS]) (Ramosaj A., 2022). The formula is the following:

$$RS = \sum_{i=1}^{n} (\hat{y}_i - y_i) \quad (15)$$

$\hat{y}_i > y_i$ yields a too optimistic forecast, whereas $\hat{y}_i < y_i$ yields a too pessimistic forecast.

3. The methodology

The methodology helped to define the most suitable model to forecast sales of accessories. The dataset was composed of 27 items with the weekly sales from 2018 to 2023 for every item. Some items were end of life (EOL), that is, the company is selling the remaining stock, but they will not order them anymore. Therefore, the three EOL items of our dataset were excluded. Of the remaining 24 items, we noticed that 11 have big accuracy impacts, whether because they are seasonal-driven, have high volatility, or for other reasons. The KAM forecasts are either too optimistic or too pessimistic.

The methodology started with the raw data, which were divided into two subsets with the monthly sales to one side and the KAM forecasts to the other. Next, the subsets were grouped by items and months to create a monthly time-series for every item. Thereafter, every sales time-series was standardized, and the computed mean (μ) and the computed standard deviation (σ) of the defined item were reused to standardize the equal item with the KAM forecasts.

Once the standardization was completed, the time-series were split into two subsets. The first contained the sales values used to fit the model, and the second was composed of the last 12 sales values, which helped to compute the metrics. The same was done with the KAM values, which were also split into data used for the fitting and data used for the computation of the metrics. This process resulted in two subsets, called the train set and test set, for the sales and KAM forecasts.

The train set of sales was used with most of the model categories (exponential smoothing, SARIMA, SARIMAX, ML, and ML with KAM) to fit the model due to make predictions for the next 12 months. These predictions were compared to the real sales of the last 12 months, also called the test set of sales, by two metrics, the RMSE and the MAE. The third metric, RS, was used at the end to compare the RS of the KAM with the best selected model. The train set of the KAM could also be used with specific model categories, such as SARIMAX and ML, with the KAM predictions. The test set of the KAM was used to measure the error between sales and KAM forecasts over the last 12 months. All RMSE

and MAE results were collected to compute the average and median for each item through the errors of the statistical and ML models. This step would define the selection criteria of a model to be able to make acceptable predictions for the time-series. The count by statistical and ML models was plotted with a histogram. The most suitable models were those that were more accurate than the KAM forecasts for each criterion, and they were ranked by largest occurrence.

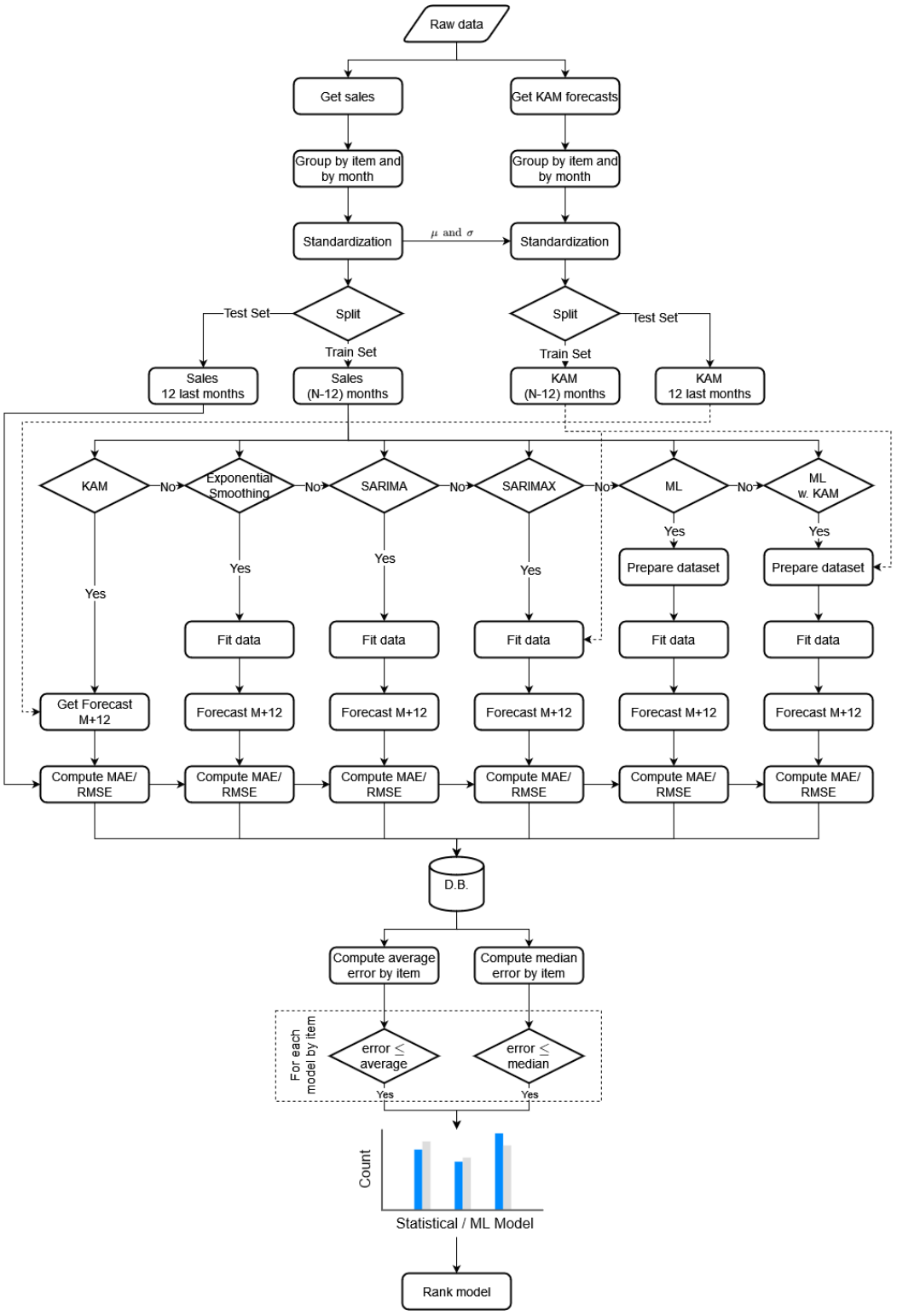

Figure 3: Methodology designed by the authors.

4. Results

The results are presented in sub-sections in order to follow the steps of the development.

The following are explanations for the differences between MAE and RMSE:
- RMSE gives more weight to large value => ability to handle outliers.
- MAE treats errors equally => unsensitivity to outliers.
- RMSE and MAE are interpretable and have same unit as the data.
- Lower RMSE and MAE indicate a better fit of the model.

The first step of the methodology consisted of obtaining the sales and KAM forecasts from the raw data and represent them by months. Figure 4 presents an example of the sales and KAM forecasts represented as a time-series.

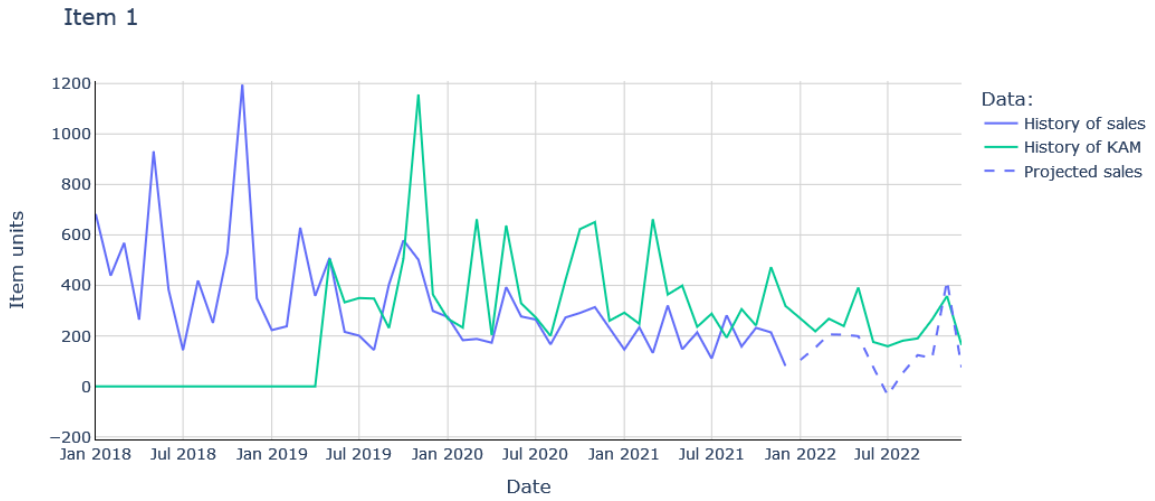

Figure 4: Presentation of the sales and KAM forecasts as a time-series. The plain lines are the history of the sales and KAM forecasts used for the training and the dashed line is the projected sales to evaluate the statistical and ML models.

Once the time-series for sales and KAM forecasts were formed, the sales values were standardized for every item individually. The mean (μ) and standard deviation (σ) from the item were also used to standardize the same item with the KAM forecasts. Figure 5 shows the standardization of the sales and KAM forecasts previously illustrated in Figure 4. The standardization allowed the distribution of the data to be retained and the mean to be brought to 0 and the standard deviation to 1. The forecasts stay in the same layout in both Figure 4 and Figure 5; the difference is that the data were processed on a common scale.

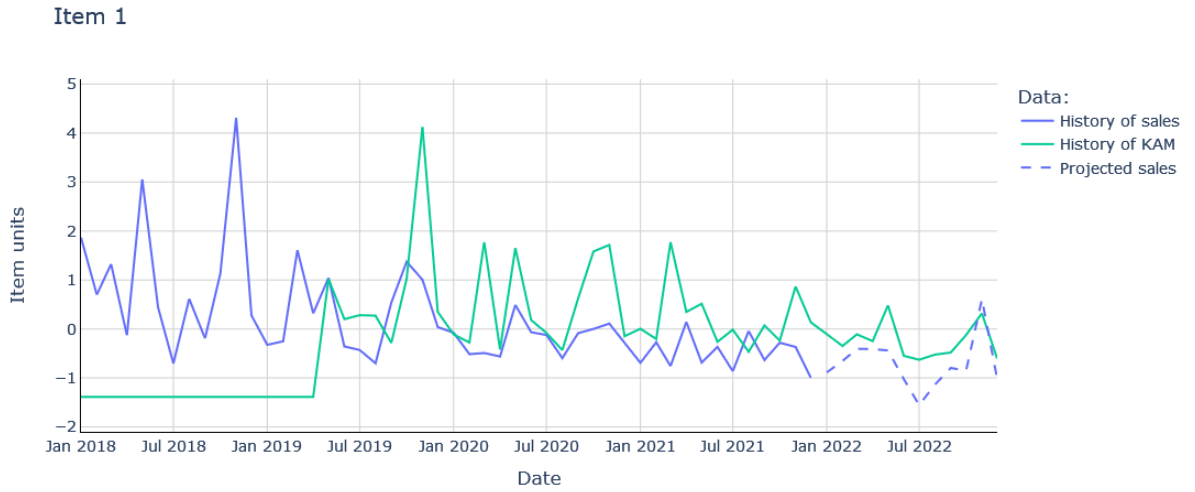

Figure 5: The sales and KAM forecasts are standardized.

Once the statistical and ML models had been performed and the error had been computed, the count of models that had an error less than or equal to the selection criteria was presented. The maximum value of counts is 24.

Based on the 16 models, it emerged that 13 models performed better than or as well as the KAM with the average criterion (Figure 6) and 14 models with the median criterion (Figure 7) with the MAE metric. Additionally, 15 models performed better than the KAM with the average criterion (Figure 8) and the median criterion (Figure 9) with the RMSE metric.

The most suitable models, which demonstrated better performances than the KAM forecasts, are the SARIMAX model coupled with the monthly mean of the KAM forecasts, the SARIMAX model associated with the last year of KAM forecasts, and the SARIMAX model without exogenous values regarding the MAE metric. As regards the RMSE metric, almost all models presented an acceptable performance except the SARIMAX model with the average of the monthly sales as exogenous values.

As shown in Figure 6, which uses the average of MAE by item as the selection criterion, nearly all models gave better predictions than the KAM forecasts, and only the exponential smoothing model and SARIMAX models with the average of the monthly sales were globally less accurate than the KAM forecasts. The model that was generally better than the KAM in most cases is the SARIMAX model with the last 12 months' KAM forecasts.

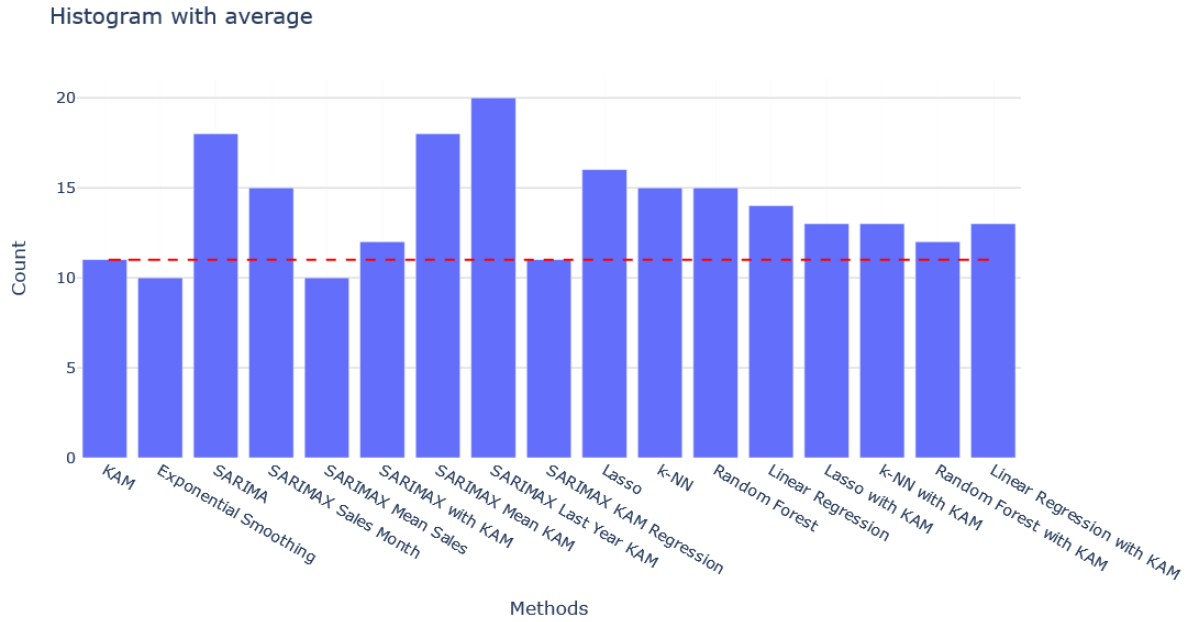

Figure 6: Histogram of the best statistical and ML models related to the KAM count (red dashed line) based on the average of the MAE by item.

Then, by using the MAE median by item as the selection criterion, as illustrated in Figure 7, almost all models presented a better or equivalent prediction than the KAM forecasts; the exceptions are the SARIMAX models with the average of the monthly sales and the SARIMAX models with the KAM forecasts fitted through a linear regression. The model that performed best is the SARIMAX model with the last 12 months of KAM forecasts.

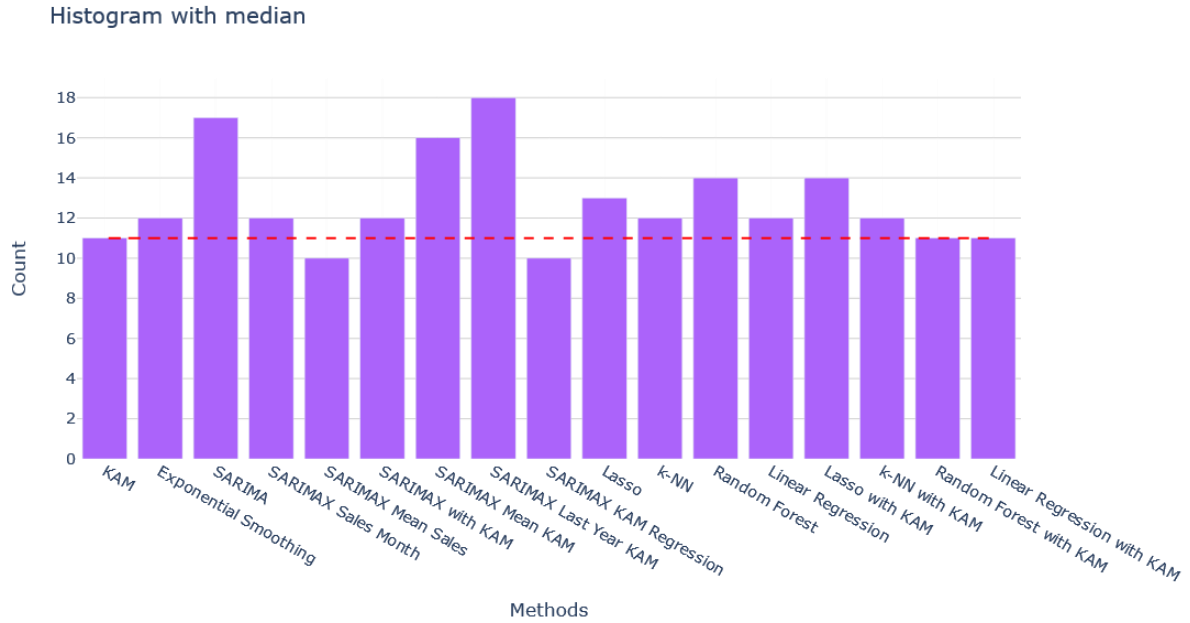

Figure 7: Histogram of the best statistical and ML models related to the KAM count (red dashed line) based on the median of the MAE by item.

Then, a ranking of the statistical and ML models based on the average and median criteria was performed and highlighted the best models by criterion. Table 3 shows the best models by metric. The three models that performed best are the SARIMAX model with the last 12 months KAM forecasts, the SARIMA model, and the SARIMAX model with the average of the monthly KAM forecasts.

Rank	Based on the average criterion		Based on the median criterion	
	Model	Count	Model	Count
1	SARIMAX last year KAM	20 / 24	SARIMAX last year KAM	18 / 24
2	SARIMA	18 / 24	SARIMA	17 / 24
3	SARIMAX mean KAM	18 / 24	SARIMAX mean KAM	16 / 24

Table 3: Ranking of the three best model by selection criteria (average, median) and their related counts with the MAE.

Then, looking at the average of RMSE by item as the selection criterion (Figure 8), almost all the models gave better forecasts than the KAM forecasts; only the SARIMAX model with the average of the monthly sales was globally less accurate than the KAM forecasts. The models that were generally better than the KAM in most cases are the SARIMAX model with the last 12 months' KAM forecasts and the SARIMAX model with the average of monthly forecasts.

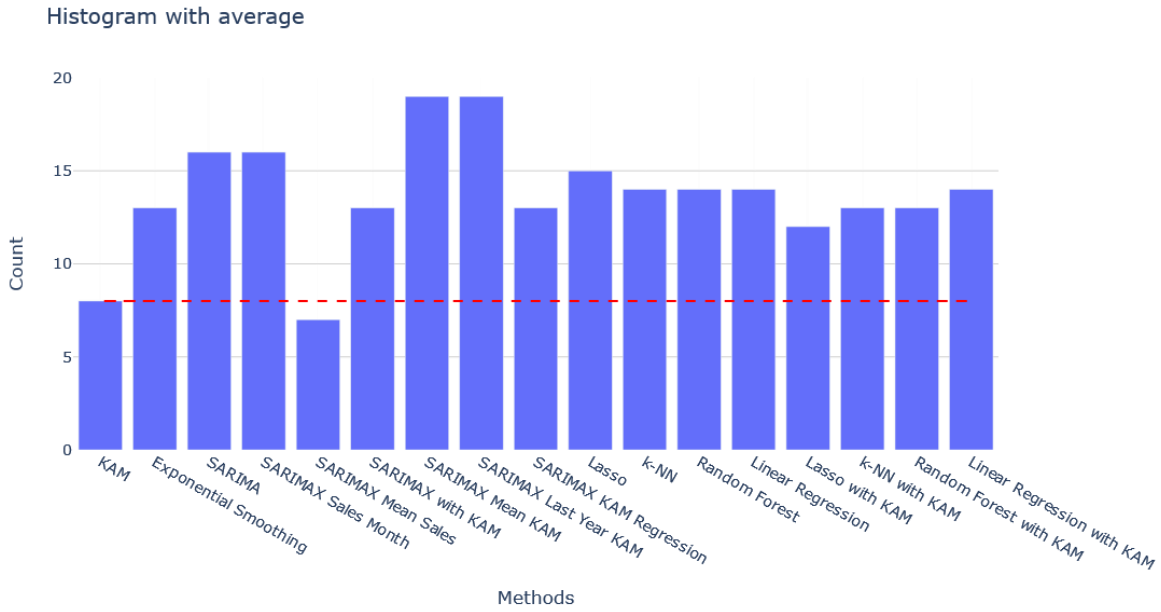

Figure 8: Histogram of the best statistical and ML models related to the KAM count (red dashed line) based on the average of the RMSE by item.

Figure 9 shows the RMSE median by item as selection criterion: all models presented a prediction that was better than or equal to the KAM. The model that performed best is the SARIMAX model with the average of the monthly KAM forecasts.

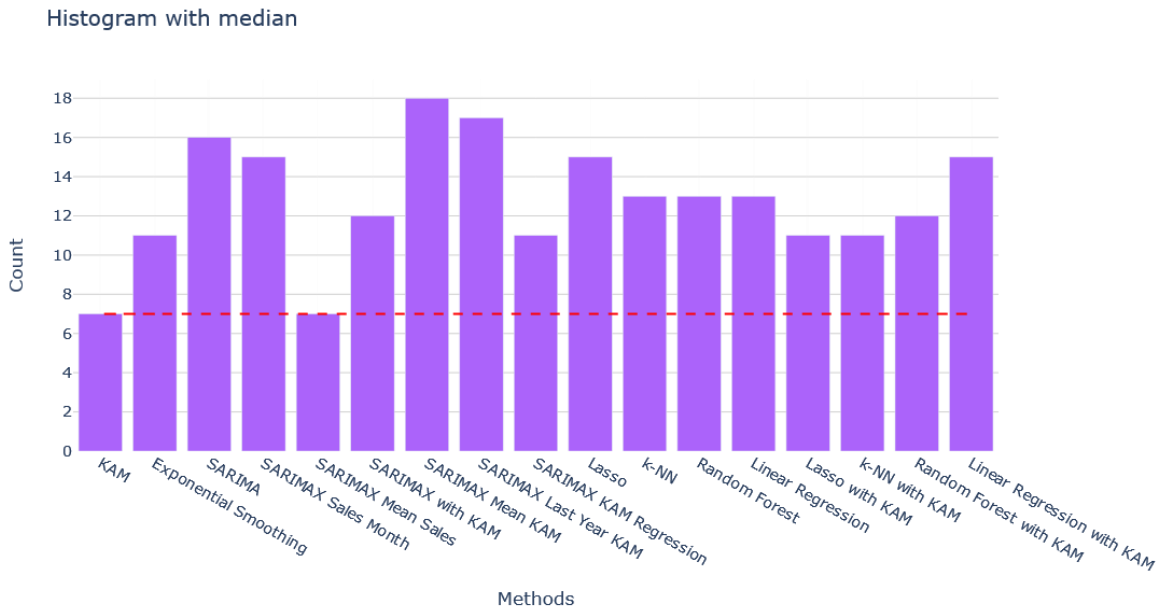

Figure 9: Histogram of the best statistical and ML models related to the KAM count (red dashed line) based on the median of the RMSE by item.

The ranking for the metric RMSE highlights three best models that are the SARIMAX model with the average of the monthly KAM forecasts, the SARIMAX model with the last 12 months KAM forecasts and the SARIMA model with the average and the median criteria. Table 4 shows the ranking and the number of occurrences.

Rank	Based on the average criterion		Based on the median criterion	
	Model	Count	Model	Count
1	SARIMAX mean KAM	18 / 24	SARIMAX mean KAM	18 / 24
2	SARIMAX last year KAM	18 / 24	SARIMAX last year KAM	17 / 24
3	SARIMA	16 / 24	SARIMA	16 / 24

Table 4: Ranking of the three best model by selection criteria (average, median) and their related counts with the RMSE.

Overall, the most suitable models to forecast the next 12 months of all items by using a single type of statistical or ML model remain the SARIMAX model with the 12 last months of KAM forecasts as exogenous values, the SARIMAX model with the average of the monthly KAM forecasts as exogenous values, and the SARIMA model. These models presented overall better forecasts than the other models based on the presented method of analysis by giving an overall reduced MAE. Thus, they presented a reduction in prediction errors and an ability to handle outliers well with a higher reduction of RMSE than the other models.

To confirm the suitability of these models, the residual stock can be used with different items presented during the test period over which the KAM forecasts were made one year ago and the last 12 month of sales were registered. A typical result is shown in Figure 10, where the red line represents the limit to avoid an OOS situation. By considering the residual stock of the KAM forecasts on 01.01.2022 without knowing the sales for the year 2022, the SME's residual stock would have increased by 1184 units, whereas the statistical model would have increased the residual stock by only 585 units. Moreover, there would

have been no stock shortages.

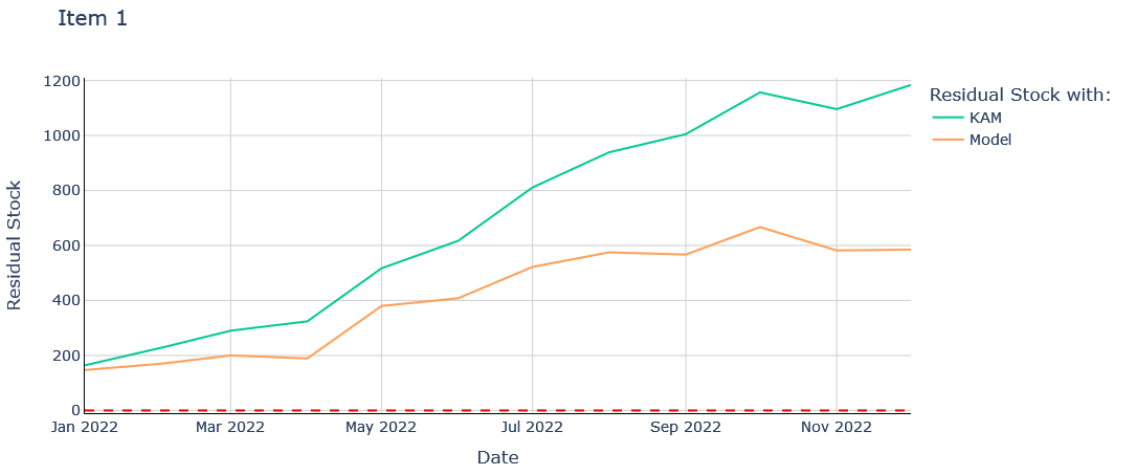

Figure 10: Residual stock for "item 1" by comparing the residual stock with the KAM forecasts vs. the statistical model SARIMAX with the 12 last months of KAM forecasts as exogenous value.

Finally, we report the residual stock after 12 months considering the KAM forecasts only or by using a statistical model, as illustrated in Figure 11. The RS show that with the 12 last months of KAM forecasts as exogenous values, the results are on 16 times over 24 times better. It is also observed that when the KAM is too extreme, the statistical model will reduce the forecasting error.

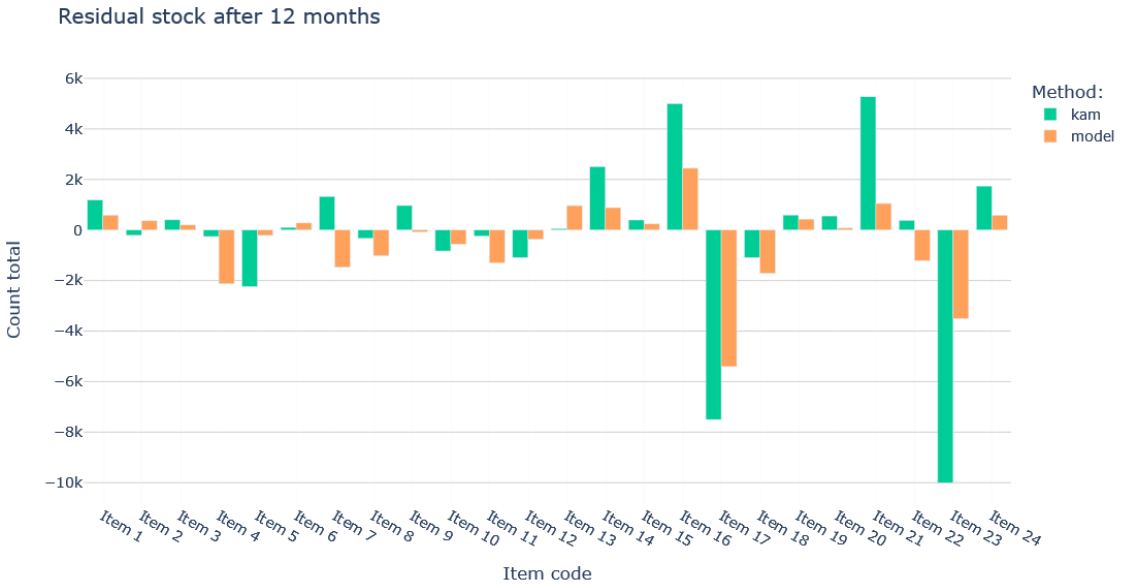

Figure 11: Residual stock for all items after 12 months when considering the KAM forecasts for the next 12 months of operation or considering the statistical model SARIMAX with the last 12 months of KAM forecasts as an exogenous value.

The same analysis can be made with the SARIMAX model with the average of the monthly forecasts of the KAM as exogenous values. This solution also presents a better residual stock for many items but tends to be more pessimistic and deliver on the long-term negative stock, posing a risk of stock shortages. Figure 12 presents the residual stock of item 1; the residual stock is kept in a range of 40 units to 318 units through the months for the statistical model opposite to the KAM forecasts that imply a continuous increasing of stock level. The red line represents the limit to avoid an OOS situation.

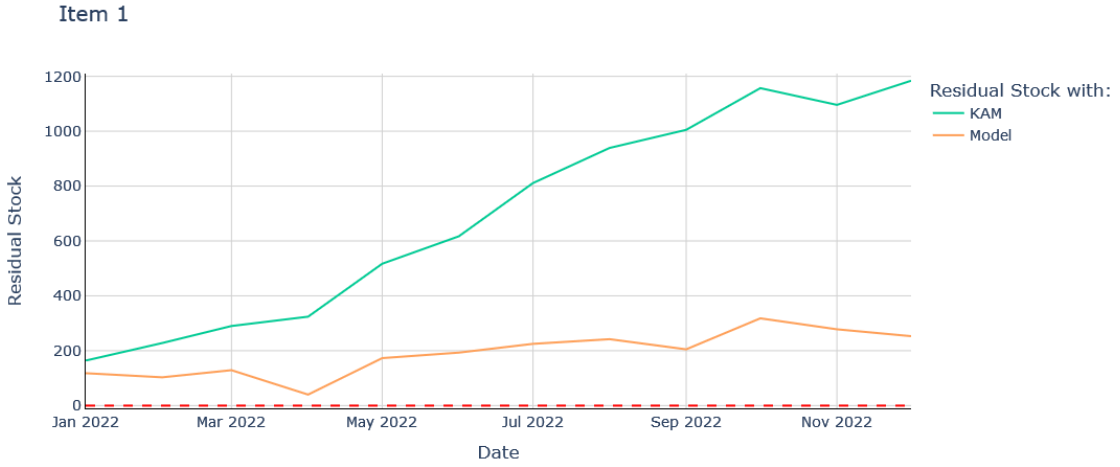

Figure 12: Residual stock for "item 1" by comparing the residual stock with the KAM forecasts vs. the statistical model SARIMAX with the average of the monthly KAM forecasts as an exogenous value.

Considering all items and looking at the residual stock after 12 months, as reported in Figure 13, it emerges that 14 of 24 forecasts made with the statistical model gave better results than the KAM forecasts. It is observed that the statistical model SARIMAX with the average of the monthly KAM forecasts increased the risk of a stock shortage.

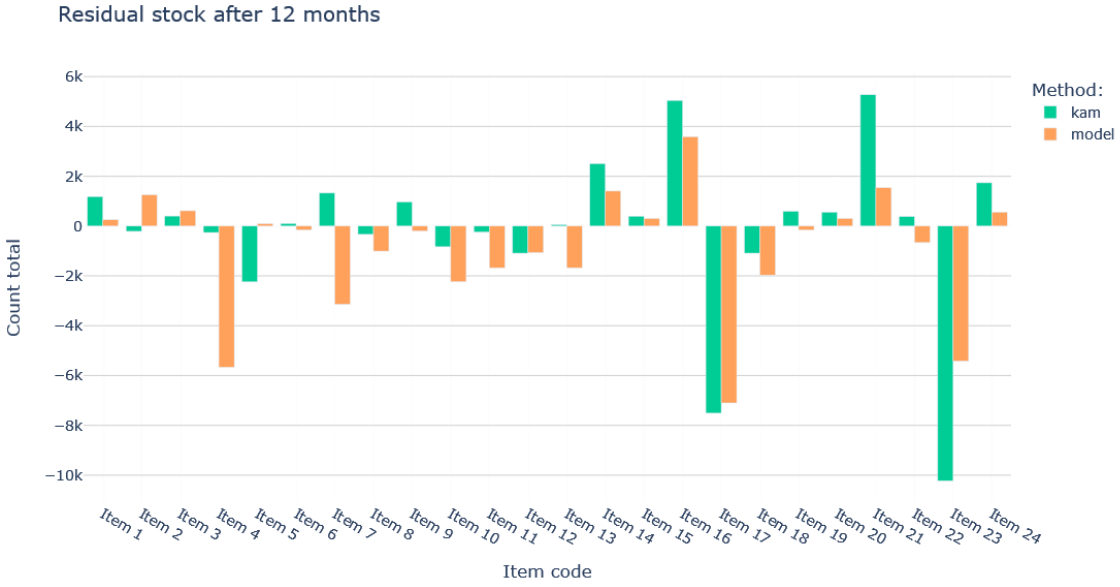

Figure 13: Residual stock for all items after 12 months when considering the KAM forecasts for the next 12 months of operation or considering the statistical model SARIMAX with the average of the monthly KAM forecasts as an exogenous value.

Finally, the analysis was made with the SARIMA model, which showed a poor residual stock because the model tended to deliver negative stock during the evaluated period. Hence, the company would risk a stock shortage. Figure 14 presents the residual stock of item 1; the residual stock is kept in a range of -92 units to 240 units through the months for the statistical model (orange), which gives good accuracy but implies a risk. The green line is the KAM forecasts, where the stock is growing continuously. The red line represents the limit to avoid an OOS situation.

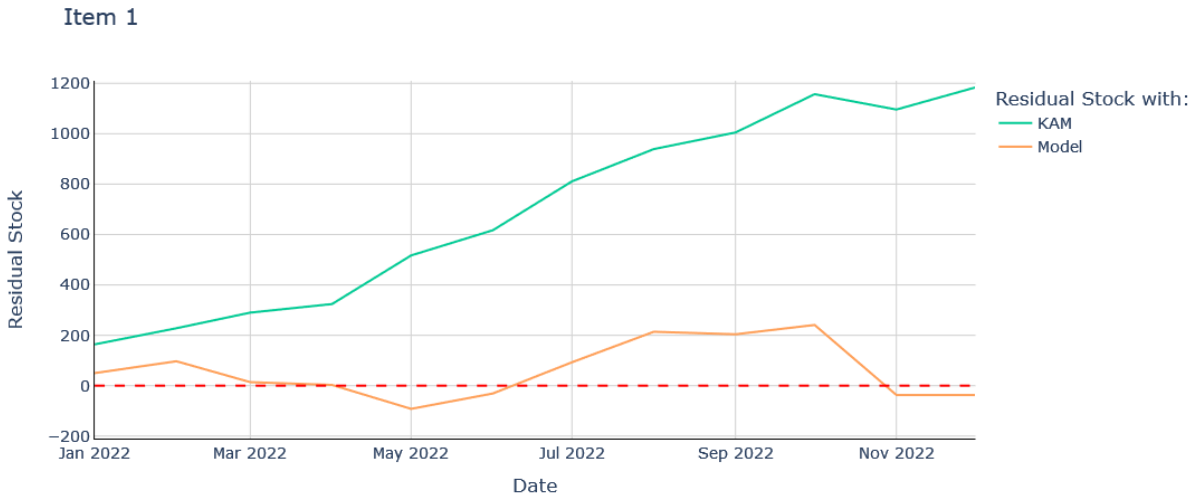

Figure 14: Residual stock for "item 1" by comparing the residual stock with the KAM forecasts vs. the statistical model SARIMA.

When all items are considered and the residual stock is examined after 12 months, as shown in Figure 15, 14 of the 24 forecasts produced using the statistical model outperformed the KAM forecasts with better accuracy. However, it is also observed that the statistical model SARIMA entailed a large risk of stock shortage.

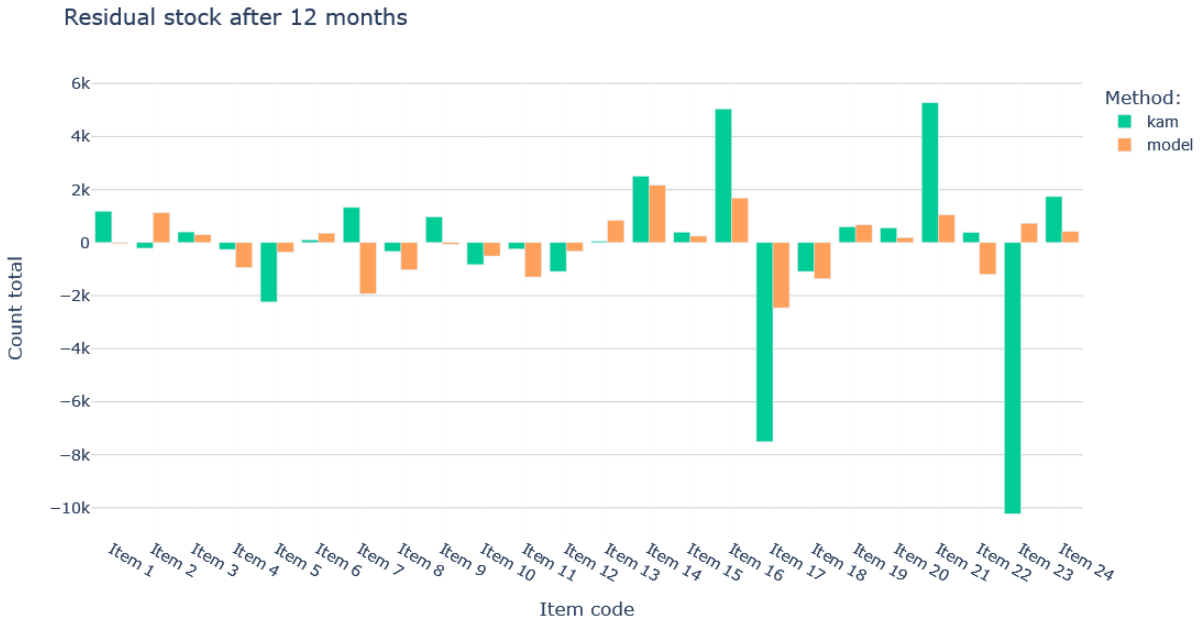

Figure 15: Residual stock for all items after 12 months when considering the KAM forecasts for the 12 next months of operation or considering the statistical model.

In general, the SARIMAX model with the last 12 months of KAM forecasts as exogenous values is the most suitable for the provided dataset of accessories. The forecasts for 2022 would be improved for 16 of 24 items, and the risk of the stock shortage would have been only present for 10 items for which the risk would have been the same if only the KAM forecasts had been considered. SARIMA and SARIMAX with the monthly average of KAM forecasts models represent a higher risk of OOS for the company. In a long-term observation, the residual stock could help to define the safety stocks needed for each item. The residual stock seems to stabilize after a few months, as shown in Figure 12.

5. Conclusion

Multiple methods were tested, and it is demonstrated that company forecasting is not completely independent of human factors. Therefore, in this paper, the human factor was combined with the best tested methods. It has been shown that statistical models perform better than ML models for accessories forecasts in general. Among the 16 models tested, 13 models performed better than the KAM. The most suitable models that performed better than the KAM forecasts are the SARIMAX model coupled with the monthly mean of the KAM forecasts, the SARIMAX model associated with the last year of KAM forecasts, and the SARIMAX model without exogenous values regarding the MAE and RMSE metrics.

Residual stock helps us to measure the remaining stock of the best selected model compared to KAM forecasts. It was noted that the KAM forecasts were always more optimistic than those of the models, which would result in stock levels being too high and therefore increase additional costs. Residual stock could also help the company to define the safety stocks, as it has been observed that the residual stock stabilizes after a few months when a statistical model is applied to the forecasts.

6. Acknowledgments

We would like to express our very great appreciation to Christoph Leuenberger for his valuable and constructive suggestions during the development of this research paper.

7. References

Arunraj N.S., Ahrens D. (2015), A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, International Journal of Production Economics 170: A, pp. 321–335. https://doi.org/10.1016/J.IJPE.2015.09.039.

Arunraj N.S., Ahrens D., Fernandes M. (2016), Application of SARIMAX model to forecast daily sales in food retail industry, International Journal of Operations Research and Information Systems, pp. 1–21, https://10.4018/IJORIS.2016040101

Arvan M., Fahimnia B., Reisi Ardali M., Siemsen E. (2019), Integrating human judgement into quantitative forecasting methods: A review, Omega, 86, pp. 237–252. https://doi.org/10.1016/j.omega.2018.07.012

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

Bakay M.S., Ağbulut U. (2021), Electricity production-based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms, Journal of Cleaner Production 285:125324285. https://doi.org/10.1016/j.jclepro.2020.125324

Berner R., Judge K., (2019) The data standardization challenge. Systemic risk in the financial sector: Ten years after the great crash, pp. 135–150. https://doi.org/10.2307/j.ctvqmp0vn.12

Grannis S.J., Xu H., Vest J.R., Kasthurirathne S., Bo N., Moscovitch B., Torkzadeh R., Rising J. (2019), Evaluating the effect of data standardization and validation on patient matching accuracy, Journal of the American Medical Informatics Association, 26(5), pp. 447–456. https://doi.org/10.1093/jamia/ocy191

Hemeimat R., Al-Qatawneh L., Arafeh M., Masoud S. (2016), Forecasting spare parts demand using statistical analysis, American Journal of Operations Research, 6, pp. 113–120. https://doi.org/10.4236/ajor.2016.62014

Hu Q., Boylan J.E., Chen H., Labib A. (2018), OR in spare parts management: A review, European Journal of Operational Research 266, pp. 395–414. https://doi.org/10.1016/j.ejor.2017.07.058

Hyndman, R.J., Athanasopoulos, G. (2018). Forecasting: Principles and practice (second edition). OTexts: Forecasting: Principles and Practice (2nd ed) (otexts.com)

James G., Witten D., Hastie T., Tibshirani R. (2021), An introduction to statistical learning: With applications in R, Springer US, New York, NY, pp. 59–80. https://doi.org/ 10.1007/978-1-0716-1418-1_3.

Jiafu R., Zongfang Z., Fang Z. (2009), The forecasting models for spare parts based on ARMA, WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, pp. 499–503, https://doi.org/10.1109/CSIE.2009.315

Makridakis S., Wheelwright S.C., Hyndman R.J. (2008), Forecasting methods and applications. USA: John Wiley and Sons.

Makridakis S., Spiliotis E, Assimakopoulosn V. (2018), Statistical and machine learning forecasting methods: Concerns and ways forward, PLoS ONE13, pp. 1–26. https://doi.org/10.1371/journal.pone.0194889

Mei J., He D., Harley R., Habetler T., Qu G. (2014), A random forest method for real-time price forecasting in New York electricity market, IEEE PES General Meeting | Conference & Exposition, National Harbor, MD, USA, pp. 1–5. https://doi.org/10.1109/PESGM.2014.6939932

Milligan G.W., Cooper M.C., (1988), A study of standardization of variables in cluster analysis. Journal of Classification 5, pp. 181–204. https://doi.org/10.1007/BF01897163

Morris, M. (2013), Forecasting challenges of the spare parts industry. Journal of Business Forecasting, 32, pp. 22─27.

Noureen S., Atique S., Roy V., Bayne S. (2019), A comparative forecasting analysis of arima model vs random forest algorithm for a case study of small-scale industrial load, International Research Journal of Engineering and Technology, 6(09), pp. 1812–1821. IRJET-V6I9281.pdf

Pinçea C., Turrini L, Meissner J. (2021), Intermittent demand forecasting for spare parts: A critical review, Omega 105, https://doi.org/10.1016/j.omega.2021.102513

Ramosaj A., Ramosaj N., Widmer M. (2022), Improving sales forecasting by combining key account managers' inputs and models such as SARIMA, LSTM, and Facebook Prophet, Journal of Applied Business and Economics 24(6). https://doi.org/10.33423/jabe.v24i6.5715

Ranstam J., Cook J.A. (2018), LASSO regression, BJS Statistical Editors, https://doi.org/10.1002/bjs.10895

Romeijnders, W., Teunter, R., Van Jaarsveld W. (2012), A two-step method for forecasting spare parts demand using information on component repairs, European Journal of Operational Research, 220, pp. 386–393.http://dx.doi.org/10.1016/j.ejor.2012.01.019

Sethi, J.K., Mittal, M. (2021), An efficient correlation based adaptive LASSO regression method for air quality index prediction, Earth Science Informatics 14, pp. 1777–1786. https://doi.org/10.1007/s12145-021-00618-1

Sheng F., Jia L. (2020), Short-term load forecasting based on SARIMAX-LSTM. 2020 5th International Conference on Power and Renewable Energy (ICPRE), IEEE, pp. 90–94. https://doi.org/ 10.1109/ICPRE51194.2020.9233117

Spiliotis E., Makridakis S., Semenoglou A.A., Assimakopoulos V. (2020), Comparison of statistical and machine learning methods for daily SKU demand forecasting, Operational Research, Volume 22, pp. 3037–3061. https://doi.org/10.1007/s12351-020-00605-2

Stadtler, H., Kilger, C. (Eds.). (2002). Supply chain management and advanced planning (pp. 71–96). Springer.

Teunter R.H., Syntetos A.A., Zied Babai M. (2011), Intermittent demand: Linking forecasting to inventory obsolescence, European Journal of Operational Research, 214, pp. 606–615. http://dx.doi.org/10.1016/j.ejor.2011.05.018

Teunter R.H., Duncan L. (2017), Forecasting intermittent demand: A comparative study, Journal of the Operational Research Society, 60(3), pp. 321–329. https://doi.org/10.1057/palgrave.jors.2602569

Tugay R., Oguducu S.G, (2020), Demand prediction using machine learning methods and stacked generalization. In Proceedings of the 6th International Conference on Data Science, Technology and Applications, pp. 216–222. SciTePress. https://doi.org/10.5220/0006431602160222

Van den Broeke, M. De Baets, S., Vereecke, A., Baecke, P., Vanderheyden, K. (2019), Judgmental forecast adjustments over different time horizons. Omega, 87, pp. 34–45. https://doi.org/10.1016/j.omega.2018.09.008

Xie M., Sandels C., Zhu K., Nordström L. (2013), A seasonal ARIMA model with exogenous variables for elspot electricity prices in Sweden, *10th International Conference on the European Energy Market (EEM)*, Stockholm, Sweden, pp. 1–4. https://doi.org/10.1109/EEM.2013.6607293