Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2019) 11(1), 25-40.

10.29137/umagd.473977

Cilt/Volume:11 Sayı/Issue:1 Ocak/January 2019

Araştırma Makalesi / Research Article

Clustering-based Sales Forecasting in a Forklift Distributor

Pratiwi Eka Puspita[1], Tülin İnkaya*[1], Mehmet Akansel[1]

[1] Uludağ University, Industrial Engineering Department, Görükle Campus, Bursa 16059, Turkey

***Başvuru/Received**: 07/06/2018* ***Kabul/Accepted**: 27/08/2018* ***Son Versiyon / Final Version:** 31/01/2019*

Abstract

Sales forecasting refers to the prediction of future demand based on past data. A vast literature on sales forecasting has accumulated due to its vital role in balancing demand and supply. Among these, data mining has emerged as a powerful tool to facilitate sales forecasting. In this study, we use data mining methods for accurate and reliable sales forecasts in a forklift distributor company. Monthly sales data for 100 different types of forklifts between 1998 and 2016 are used. The proposed forecasting methodology includes three steps. First, products with similar sales patterns are determined using hierarchical clustering. Dynamic time warping is applied to calculate the similarities among product sales data. Second, features are extracted and selected for each cluster. In addition to the features adopted from the literature, four new features are proposed to characterize intermittency. Multivariate adaptive regression splines model is used for feature selection. Third, support vector regression is used to predict future sales of each product cluster. Finally, the performance of the proposed approach is evaluated according to forecasting error and complexity. The numerical analysis shows that the proposed approach gives reasonable accuracy with less complexity.

Key Words

"Data mining; clustering; forecasting; multivariate adaptive regression splines (MARS); dynamic time warping (DTW); support vector regression (SVR)"

***Sorumlu Yazar:** tinkaya@uludag.edu.tr*

1. INTRODUCTION

Sales forecasting refers to the prediction of future demand based on past data. It has a vital role in today's business environment as accurate and reliable forecasts form the fundamental basis for the capacity and production planning decisions in a company. Additionally, effective forecasting helps the companies in the supply chain improve overstock and understock costs, and decreases the safety stock requirements.

A vast literature on sales forecasting has accumulated due to its vital role in balancing demand and supply. Some widely known methods for sales forecasting are moving average (MA), single exponential smoothing (Brown, 1959), Holt-Winters model (Winters, 1960), and auto regression integrated moving average (ARIMA) (Box and Jenkins, 2015). However, these traditional methods may fail when there exist noise, outliers, and intermittency in the data.

In the recent years, data mining has emerged as a powerful tool to facilitate sales forecasting to overcome the limitations of the traditional forecasting approaches. Data mining refers to extracting the interesting, non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from large amounts of data (Han et al., 2012). Data mining tasks are classified into two categories: descriptive and predictive. Descriptive tasks, such as clustering and association rule mining, characterize the general properties of the data, whereas predictive ones, such as classification and regression, make predictions using the existing data (Han et al., 2012). In this study, we consider sales forecasting as a prediction problem for time series data.

A number of studies suggest that support vector regression (SVR) has gained considerably wider acceptance in time series forecasting, including intermittent data (Bao et al., 2005), due to its strengths compared to other approaches (Levis and Papageorgiou, 2005; Yu et al., 2013). Nalbantov et al. (2005) claim that SVR can be used to avoid overfitting problems and to improve the robustness of outlier detection. In addition, Thissen et al. (2003) explain that SVR implementation has advantages, such as finding a globally optimal solution and calculating a nonlinear solution efficiently. Das and Padhy (2012) discuss the advantage of SVR compared to the use of back propagation neural network (BPN) in forecasting the non-linear time series of stock market. Zuo et al. (2014) obtain the best outcome with SVR model compared to linear discriminant analysis, logistic regression, and Bayesian network for the Radio Frequency Identification (RFID) data of consumer in-store behavior.

Hybridization of SVR with other methods improves the forecasting accuracy. Wisner et al. (2015) state that integrated forecasting is expected to reduce errors. Hua and Zhang (2006) conclude that the hybridization of logistic regression and SVR (LRSVM) outperforms the forecasting methods for intermittent time series such as Croston's method (Croston, 1972), Markov bootstrapping, and single SVR.

Some studies focus on the selection of the useful features for SVR. Lu et al. (2009) use independent component analysis (ICA) to remove the features containing noise. ICA together with SVR results in better accuracy in forecasting financial time series compared to pure SVR. Also, Lu et al. (2012) perform feature selection with multivariate adaptive regression splines (MARS). In a recent study, Lu (2014) extracts features adopted from the technical indicators of the stock market and characterizes different properties of the data set, i.e. trend, growth ratios, and sales volatility.

A stream of studies conducts customer segmentation before forecasting so that customized forecasting models are developed for customers with similar characteristics. For this purpose, clustering methods are adopted for customer segmentation such as hierarchical clustering (Huber et al., 2017; Biscarri et al., 2017), k-means (Kuo and Li, 2016; Dai et al., 2015) and fuzzy c-means (Bao et al., 2004). Clustering can be performed using categorical variables such as customers' properties (Biscarri et al., 2017) or time series data (Lu and Kao, 2016; Chen and Lu, 2017). For example, Bala (2012) first clusters the customers according to their demographic data and purchasing behavior. Then, for each customer cluster (segment), an autoregressive integrated moving average (ARIMA) is used for forecasting. Dai et al. (2015) apply k-means algorithm to extract the disjoint clusters in the aggregate sales of computer servers. Then, for each cluster, SVR is used for forecasting.

In most of these studies, Euclidean distance is used for calculating the dissimilarities among the products. However, products may have different release and phase-out times, so the lengths of the sales data may differ. In this case, it is not convenient to use Euclidean distance as a dissimilarity measure. Dynamic time warping (DTW) distance (Berndt and Clifford, 1994) is able to calculate the distances between temporal sequences having different lengths. It provides the optimal alignment between two temporal sequences. Murray et al. (2017) use DTW distance for customer segmentation, however, they only focus on time series clustering. Different from their study, we combine clustering, feature selection and extraction, and forecasting, and by doing so propose an integrated forecasting methodology.

In this study, we consider sales forecasting problem for a forklift distributor. Monthly sales data for 100 different types of forklifts between 1998 and 2016 are used. A new clustering-based forecasting methodology is proposed. The proposed approach includes three steps. In the first step, the products having similar sales patterns are determined using hierarchical clustering. Products have different release and phase-out times; therefore, the lengths of the sales data are not equal. For this purpose, we adopt dynamic time warping (DTW) as a distance measure in clustering-based forecasting. We also determine the representatives of each cluster. In the second step, feature extraction and selection are performed using MARS. In addition to the features used for time series data, new features are proposed for intermittent data. In the third step, SVR is used for forecasting. The proposed approach aims to decrease the number of forecasting models via clustering.

As a summary, the contributions of this paper are as follows:

1. A new forecasting methodology based on data mining is proposed. The proposed methodology integrates the clustering, feature extraction, feature selection, and prediction tasks of data mining.
2. DTW is adopted as a distance measure for clustering sales data with unequal lengths.
3. New features are proposed for forecasting intermittent data.

The rest of the paper is organized as follows. Section 2 explains the methods used in the study. Section 3 introduces the proposed clustering-based forecasting methodology. Section 4 includes the numerical studies and results. Finally, Section 5 provides the conclusion and directions of future work.

2. METHODS

This section explains the methods used throughout the paper.

2.1. Clustering

Clustering is the process of grouping a set of objects such that similar objects are in the same cluster and dissimilar objects are in different clusters (Han et al., 2012). It has applications in various fields such as market segmentation, image segmentation, pattern recognition, and so on. Basically, the clustering methods are classified into five categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and probability-based methods.

In this study, we use the hierarchical clustering, which groups data points into a tree of clusters (Han et al., 2012). According to the hierarchical decomposition method, there are agglomerative (bottom-up) and divisive (top-down) approaches. In the agglomerative approach, each cluster is initialized as a data object, and then clusters are merged until all objects are in a single cluster. The divisive version does the reverse of the agglomerative version.

The linkage scheme defines how the distance between the clusters is calculated. Single-linkage clustering uses the minimum distance to determine the distance between two clusters. When the maximum distance is used to calculate the distance between two clusters, it is known as complete linkage. Average linkage uses the average of the distances between two clusters. Ward's linkage algorithm considers the sum of squares between clusters (Han et al. 2012). Also, hierarchical clustering can be used with various distance measures including DTW.

2.1.1. Cluster representative

In time series clustering, cluster medoid is commonly used as the cluster representative (Hautamaki et al., 2008). That is, the data object having the minimum total distance to the rest of the cluster members is selected as the representative:

$$H_i = \arg\min_{S_j \in C_i} \left\{ \sum_{S_k \in C_i \setminus S_j} d(S_k, S_j) \right\} \tag{1}$$

where H_i is the representative for cluster i, d is the distance measure, S_k is data object k, and C_i is the set of data objects in cluster i.

Another method for finding the cluster representative is DTW Barycenter Averaging (DBA) (Petitjean et al., 2011). This approach minimizes the sum of squared DTW distances from the average sequence, namely barycenter, to the sequences of all time series in the cluster. Technically, let $\mathbb{S} = \{S_1,..,S_N\}$ be the sequences of time series in cluster, and $C = \left\langle C_1,...,C_T \right\rangle$ be the average sequence of $\mathbb{S}$ at iteration *i*. DBA minimizes the within group sum of squares (WGSS) iteratively as follows:

$$\mathrm{WGSS}(C) = \sum_{k=1}^{N} d_{DTW}^2(C, S_k) \tag{2}$$

In each iteration, two steps are performed: 1) DTW distance between the average sequence and each sequence in the cluster is computed, and 2) each coordinate in the average sequence is updated as the barycenter of the coordinates associated with it.

2.1.2. Dissimilarity measure

In clustering, the dissimilarities among the objects are measured using various distance functions. Euclidean distance is often used to calculate the dissimilarity between two data objects (Agrawal et al., 1993). The Euclidean distance between vectors X_i and X_j is calculated as follows:

$$d\left(X_i, X_j\right) = \sqrt{\sum_{k=1}^{n}\left(X_{ik} - X_{jk}\right)^2} \tag{3}$$

where X_{ik} and X_{jk} denote the *k*th attributes of vectors X_i and X_j respectively, and *n* is the number of attributes.

Although it is used in several studies on customer segmentation (Thomassey and Fiordaliso, 2006; Kumar and Rathi, 2011; Chen and Lu, 2017), it has limitations in datasets with unequal lengths (Keogh, 1997).

As a remedy, dynamic time warping (DTW), which is an elastic measure, is introduced (Berndt and Clifford, 1994). DTW calculates the dissimilarity between two sequences of time series with unequal lengths. Let $Q = (q_1, .., q_n)$ and $P = (p_1, .., p_m)$ be two sequences with lengths *n* and *m*. A *n*-by-*m* matrix is constructed such that the $(i, j)^{\text{th}}$ element shows the Euclidean distance between $d(q_i, p_j)$. A warping path, *W*, depicts a mapping between *Q* and *P*, and the *k*th element of W is defined as $w_k = (i, j)_k$. So, the warping path becomes:

$$W = w_1,..,w_K \quad \max\left(m,n\right) \le K \le m + n - 1 \tag{4}$$

The warping path is subject to constraints, i.e. boundary conditions, continuity, and monotonicity. Boundary conditions require the path to start from $w_1 = (1, 1)$ and to finish at $w_K = (m, n)$ in the diagonally opposite corner of the matrix. For continuity, the allowable steps are restricted, i.e. given $w_K = (a, b)$ then $w_{k-1} = (a', b')$ where $a - a' \le 1$ and $b - b' \le 1$. Also, the points in *W* are forced to be monotonical, given $w_k = (a, b)$ then $w_{k-1} = (a', b')$ where $a - a' \ge 0$ and $b - b' \ge 0$.

The aim is to find the warping path with the minimum warping cost:

$$DTW\left(Q,P\right) = \min\left\{\sum_{k=1}^{K} w_k\right\} \tag{5}$$

The optimal path can be calculated using dynamic programming to assess the following recursive function:

$$\gamma(i,j) = d\left(q_i, p_j\right) + \min\left\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\right\} \tag{6}$$

where $\gamma(i,j)$ is the cumulative distance between points q_i and p_j.

2.2. Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) is a nonparametric regression procedure to model the interactions between dependent and independent variables without any assumption about their functional relationship (Friedman, 1991). It can handle data sets with high-dimensionality. Besides, MARS can investigate the important variables without long training processes, and saves computation time (Lu et al., 2012).

MARS uses the so-called basis function (*t-x*) and (*x-t*), where *t* is the knot of the basis functions, to approximate the linear or nonlinear relationships. Only positive part of the basis functions is considered, otherwise it takes a value of zero. The technique starts with the simplest model of the basis function. Then, it continues with adding the basis function (for each variable and for all possible knots) recursively so that prediction error is minimized.

The general MARS function can be defined as follows (Lu, 2014):

$$f(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[S_{km} \left(x(k,m) - t_{km} \right) \right] \tag{7}$$

where a_0 is intercept, a_m is the coefficient of the model, *M* is the number of basis functions, K_m is the number of knots, S_{km} is the right/left position of the associated step function, $x(k, m)$ is the label of the independent variable, and t_{km} is the knot location.

The details of MARS are provided by Friedman (1991).

2.3. Support Vector Regression

Support vector regression (SVR) (Vapnik, 1995) is based on statistical learning theory, and it is a version of support vector machine (SVM) for regression. SVR can be formulated as follows (Vapnik, 1995):

$$f(x) = \left(w \cdot \phi(x) \right) + b \tag{8}$$

where *w* is the weight vector, *x* is the model input, $\phi(x)$ is the kernel function to transform the non-linear inputs to linear form, and *b* is the bias.

Given training data $\{(x_1, y_1),..,(x_n, y_n)\} \subset \mathbb{R}$, the aim is to find a function *f*(*x*) that deviates at most ε from the target values in the training data. The slack variables ξ_i and ξ_i^* are introduced to allow errors beyond ε precision. Hence, the weight vector (*w*) and bias (*b*) are estimated by using the following convex optimization problem:

Minimize,

$$z = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{9}$$

Subject to

$$y_i - \left(w \cdot \phi(x_i) \right) - b \le \varepsilon + \xi_i \qquad i = 1,...,n \tag{10}$$

$$\left(w \cdot \phi(x_i) \right) + b - y_i \le \varepsilon + \xi_i^* \qquad i = 1,...,n \tag{11}$$

$$\xi_i, \xi_i^* \ge 0 \qquad i = 1,...,n \tag{12}$$

where $C > 0$ is a constant to specify the trade-off between $\|w\|^2$ (flatness of function *f*) and the tolerance to deviations larger than ε.

Using Lagrangian multipliers and Karush-Kuhn-Tucker conditions, Equations (9)-(12) transform into the dual Lagrangian form as follows:

Maximize,

$$L_d\left(\alpha,\alpha^*\right) = -\varepsilon\sum_{i=1}^{n}\left(\alpha_i + \alpha_i^*\right) + \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)y_i - \frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)K\left(x_i, x_j\right) \tag{13}$$

Subject to

$$\sum_{i=1}^{n}\left(\alpha_i^* - \alpha_i\right) = 0 \tag{14}$$

$$0 \le \alpha_i \le C \qquad i = 1,\dots,n \tag{15}$$

$$0 \le \alpha_i^* \le C \qquad i = 1,\dots,n \tag{16}$$

where α_i and α_i^* are the Lagrangian multipliers that satisfy $\alpha_i\alpha_i^* = 0$, and $K\left(x_i, x_j\right)$ is a Kernel function. The optimal solution of the weight vector can be calculated as $w^* = \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)K\left(x, x_i\right)$. Thus, the general form of SVR becomes as follows:

$$f\left(x, w\right) = f\left(x, \alpha, \alpha^*\right) = \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)K\left(x, x_i\right) + b \tag{17}$$

3. PROPOSED METHODOLOGY

In this section, clustering-based forecasting methodology is introduced for time series data with unequal length and intermittency. The proposed approach addresses the forecasting problem of a company with a high product variety. Therefore, the aim is to achieve high forecasting accuracy with less complexity.

The proposed methodology has three steps, and its flowchart is provided in Fig. 1. In the first step, agglomerative hierarchical clustering is used to determine the products with similar sales patterns using DTW distance. The number of clusters is determined according to the inter-cluster heterogeneity and intra-cluster homogeneity. Also, for each cluster, the cluster representative is found calculating both cluster's medoid and DBA. In the second step, feature extraction and selection are performed. Table 1 presents the 26 features adopted from Lu (2014). They characterize amount, trend, growth, and volatility. In addition to them, Table 2 shows the four new features, which we introduce to characterize the intermittency. After feature extraction, MARS is used to select the useful features for each cluster. In the third step, SVR is used to build a forecasting model for each cluster representative.

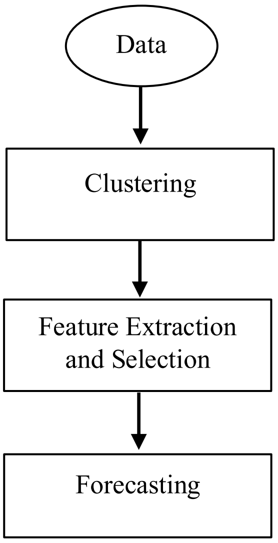


Fig. 1. Flowchart of the proposed methodology.

Table 1. Features for time series data (Lu, 2014).

Feature	Description	Period	Characteristic
$T1$	$X_1 = C_{(t-1)}$	Short term	Amount
$T2$	$X_2 = C_{(t-2)}$	Short term	Amount
$T3$	$X_3 = C_{(t-3)}$	Short term	Amount
$T5$	$X_4 = C_{(t-5)}$	Mid term	Amount
$T10$	$X_5 = C_{(t-10)}$	Mid term	Amount
$T15$	$X_6 = C_{(t-15)}$	Long term	Amount
$T20$	$X_7 = C_{(t-20)}$	Long term	Amount
$MA2$	$X_8 = \sum_{i=1}^{2} C_{(t-i)} \Big/ 2$	Short term	Trend
$MA3$	$X_9 = \sum_{i=1}^{3} C_{(t-i)} \Big/ 3$	Short term	Trend
$MA5$	$X_{10} = \sum_{i=1}^{5} C_{(t-i)} \Big/ 5$	Mid term	Trend
$MA10$	$X_{11} = \sum_{i=1}^{10} C_{(t-i)} \Big/ 10$	Mid term	Trend
$MA15$	$X_{12} = \sum_{i=1}^{15} C_{(t-i)} \Big/ 15$	Long term	Trend
$RDP1$	$X_{13} = 100 \times \left(C_t - C_{(t-1)}\right) / C_{(t-1)}$	Short term	Growth ratios
$RDP3$	$X_{14} = 100 \times \left(C_t - C_{(t-3)}\right) / C_{(t-3)}$	Short term	Growth ratios
$RDP5$	$X_{15} = 100 \times \left(C_t - C_{(t-5)}\right) / C_{(t-5)}$	Mid term	Growth ratios
$RDP10$	$X_{16} = 100 \times \left(C_t - C_{(t-10)}\right) / C_{(t-10)}$	Mid term	Growth ratios
$RDP15$	$X_{17} = 100 \times \left(C_t - C_{(t-15)}\right) / C_{(t-15)}$	Long term	Growth ratios
$BIAS5$	$X_{18} = 100 \times \left(C_t - MA5\right) / MA5$	Mid term	Volatility
$BIAS10$	$X_{19} = 100 \times \left(C_t - MA10\right) / MA10$	Mid term	Volatility
$BIAS15$	$X_{20} = 100 \times \left(C_t - MA15\right) / MA15$	Long term	Volatility
$ROC5$	$X_{21} = 100 \times C_t / C_{(t-5)}$	Mid term	Volatility
$ROC10$	$X_{22} = 100 \times C_t / C_{(t-10)}$	Mid term	Volatility
$ROC15$	$X_{23} = 100 \times C_t / C_{(t-15)}$	Long term	Volatility

Table 1. (Cont.) Features for time series data (Lu, 2014).

Feature	Description	Period	Characteristic
$Disparity5$	$X_{24} = 100 \times C_t / MA5$	Mid term	Volatility
$Disparity10$	$X_{25} = 100 \times C_t / MA10$	Mid term	Volatility
$OSCP5$	$X_{26} = 100 \times (MA5 - MA10) / MA5$	Mid term	Volatility

Note: C_t denotes the amount of sales in period *t*.

Table 2. Proposed intermittency features.

Feature	Description	Period	Characteristic
IML	$X_{27} = \sum_{k=1}^{t-1} I_k \Big/ (t-1)$	Long term	Intermittency
IMM	$X_{28} = \left(\lvert CC_{(t-1)} \rvert + \lvert CC_{(t-2)} \rvert \right) / 2$	Mid term	Intermittency
$IMS1$	$X_{29} = \lvert CC_{(t-1)} \rvert \Big/ \sum_{k \in \{k' : C_{k'} \in CC_{t-1}\}} I_k$	Short term	Intermittency
$IMS2$	$X_{30} = \sum_{k \in \{k' : C_{k'} \in CP_{t-1}\}} I_k \Big/ \lvert CP_{(t-1)} \rvert$	Short term	Intermittency

Notes: 1. I_k is an indicator variable such that $I_k = 1$ if $C_k = 0$, i.e. the amount sales in period *k* is 0.

2. CC_t is a sequence $CC_t = \left(C_{t-k}, ..., C_t \right)$ such that $C_{t-k} > 0$ and $\sum_{k'=0}^{k+1} C_{t-k'} = 0$.

3. CP_t is a sequence $CP_t = \left(C'_{t-k}, ..., C'_t \right)$ such that $C'_{t-k-1} > 0$ and $\sum_{k'=1}^{k} C'_{t-k'} = 0$.

4. CASE STUDY

The proposed forecasting methodology is applied to a distribution company for industrial heavy equipment, i.e. forklifts. The company is located in Indonesia with 15 branch offices throughout the country, and it has the biggest market share for forklift distribution in the last five years. It employs 750 workers to satisfy the demand of about 5,000 customers.

4.1. Sales Data

The monthly sales data for 100 different types of forklifts were collected from the company. The sales data span a time horizon of 19 years, from January 1998 to December 2016.

Data preprocessing operations were performed on the dataset. In this context, two products without any demand in this time interval were removed. Hence, 98 products are considered during the numerical study. Next, sales data of each product were cropped according to the release and phase-out times. These preprocessing steps yielded time series data with varying lengths between 12 and 228. Fig. 2 shows the sales data of four example products.

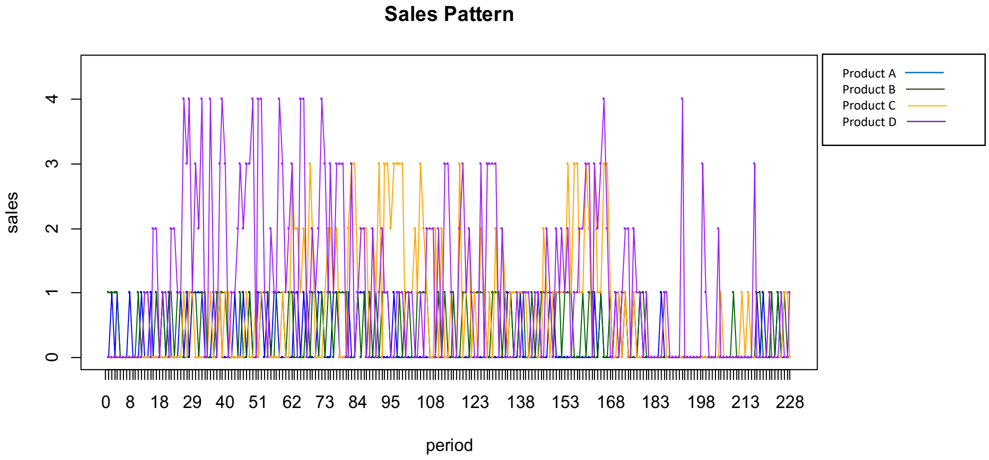


Fig. 2. Sales pattern of four example products.

Also, the intermittency level of each product is evaluated. The intermittency level for item *i*, *ilevel*(*i*), is calculated as follows:

$$ilevel(i) = \frac{\text{\# of zero demand values in item } i}{\text{length of sales data for item } i} \tag{18}$$

Fig. 3 presents the histogram of the intermittency levels for all products. The histogram indicates that more than half of the products show high intermittency, i.e. the intermittency level is higher than 0.5. Only two products have no intermittency.

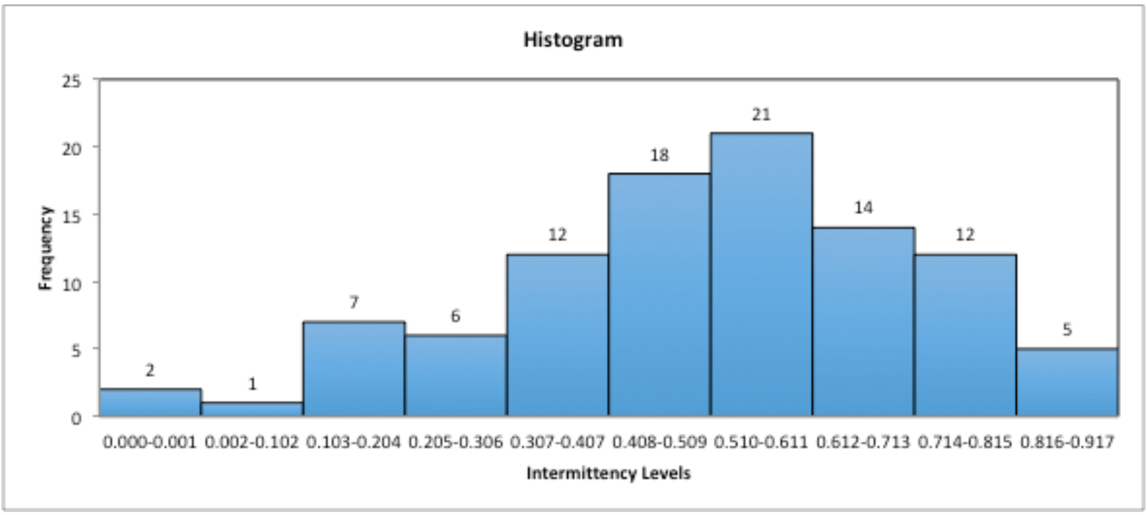


Fig. 3. Histogram of the intermittency levels for all products.

4.2. Parameter Settings and Performance Criteria

In the numerical study, agglomerative hierarchical clustering with complete linkage is used. The parameters of DTW are set as follows: i) step pattern is symmetric2, ii) window type is slanted-band, and iii) window size is 16.

MARS and SVR are evaluated using “leave-one-out” cross-validation. In SVR, radial basis function (RBF) kernel is selected, and a grid search is performed to determine the best parameter setting for C and ε. For this purpose, parameters of C and ε are iterated within a range of $[2^0, 2^{15}]$ and [0, 1], respectively.

The performance of the forecasting results is evaluated using root mean square error (RMSE), mean square error (MSE) and mean absolute deviation (MAD) (Lu, 2014). The formulas of the performance measures are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2} \tag{19}$$

$$MSE = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2 \tag{20}$$

$$MAD = \frac{\sum_{k=1}^{n}\left|y_k - \hat{y}_k\right|}{n} \tag{21}$$

where n denotes the number of periods, y_k is the actual demand for the k^{th} period, and $\hat{y}_k$ is the forecasted demand for the k^{th} period.

4.3. Numerical Results

The results of the proposed methodology are explained in this section. Note that all numerical experiments were conducted in R (R Core Team, 2017).

4.3.1. Clustering results

Determination of the number of clusters is a challenging task, as there is not a widely accepted method in the literature (Jain, 2010). Since the aim of this study is to obtain homogenous clusters with products having similar sales pattern, clustering results are evaluated according to both homogeneity and heterogeneity measures. Homogeneity is calculated as the mean pairwise DTW distance within the same cluster, whereas heterogeneity is calculated as the mean pairwise DTW distance between two clusters. While the value of heterogeneity is a "larger-the-better" measure, the value of homogeneity is regarded with a "smaller-the-better" approach.

We varied the number of clusters (k) between 2 and 30, and applied agglomerative clustering algorithm with DTW distance. The heterogeneity and homogeneity measures are plotted with respect to the number of clusters as in Fig. 4. The result indicates that homogeneity and heterogeneity measures stabilize for k = 7, 16, 27. As an example, the dendrogram for k = 7 is provided in Fig. 5. Note that, in the clustering result, there are four singletons, i.e. clusters with a single product. Also, clusters 1, 2 and 6 have 70, 21 and 3 members, respectively.

After determination of the number of clusters, the members and the representative of each cluster are obtained as shown in Fig. 6.

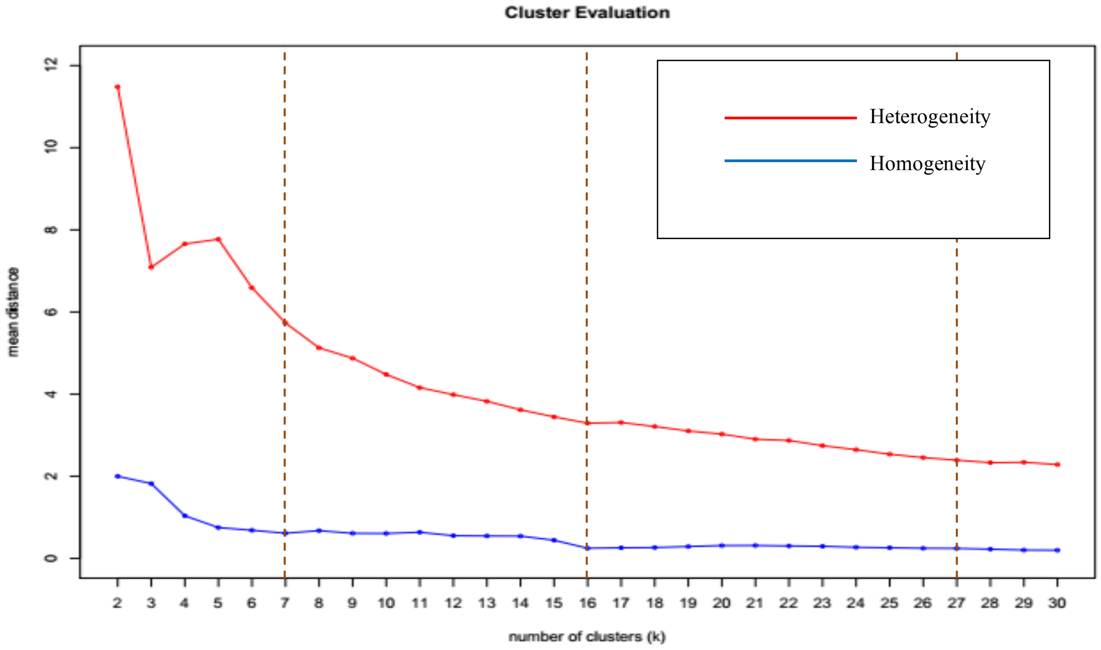


Fig. 4. Evaluation of the number of clusters with respect to homogeneity and heterogeneity measures

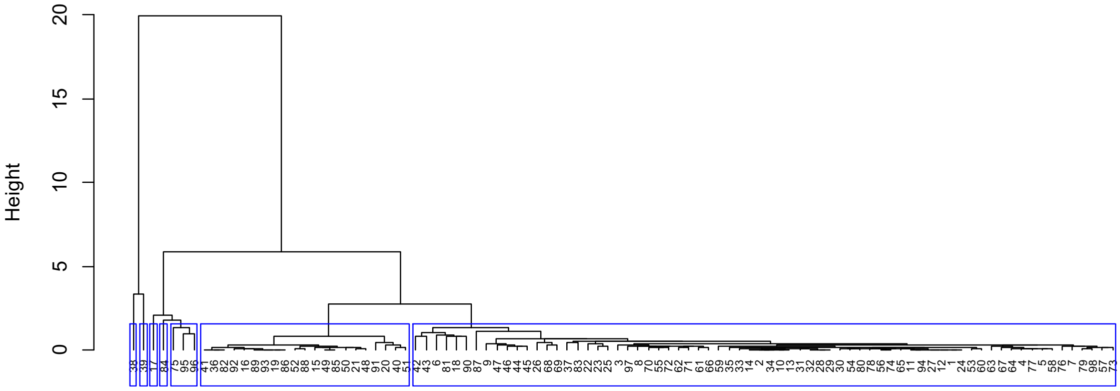
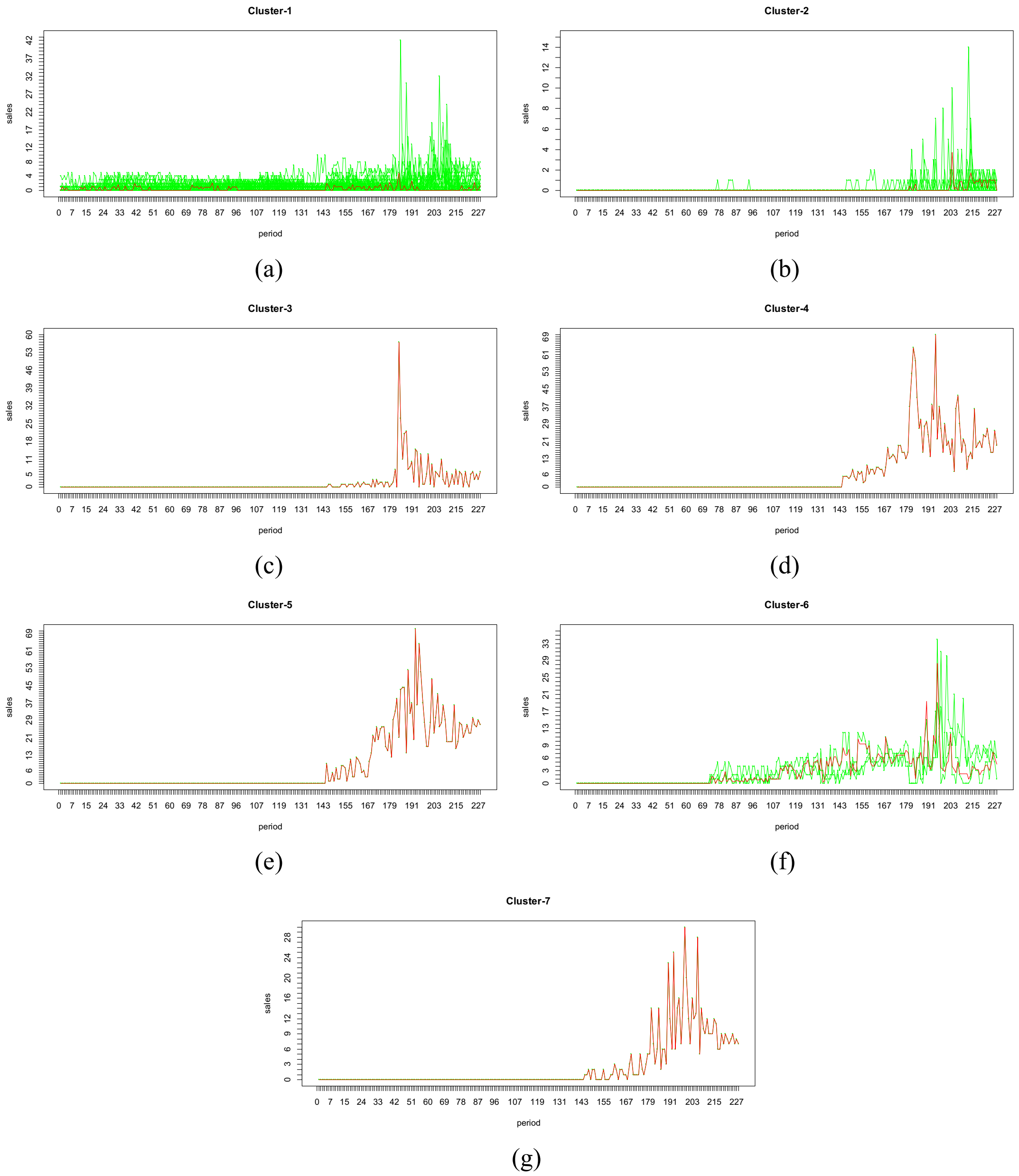


Fig. 5. Dendrogram for *k*=7 (blue rectangles show the seven clusters).



(a) (b) (c) (d) (e) (f) (g)

Fig. 6. Cluster members for *k*=7 and cluster representatives with DBA (red lines show the cluster representatives)

4.3.2. Feature extraction and selection results

The forecasting performances with and without intermittency features are compared in order to evaluate the contribution of the proposed intermittency features in Table 2. In the forecasting without intermittency features (IF), only features in Table 1 are used as input to the SVR model, whereas, in the forecasting with IF, features in both Tables 1 and 2 are used in the SVR.

The summary of the results for the products with positive intermittency levels is presented in Table 3. Table 3 indicates that RMSE, MSE, and MAD values for the SVR with IF are less than the ones without IF. Hence, the proposed features are able to characterize the intermittency of time series, and improve the forecasting errors.

After feature extraction, MARS is applied to select the useful features. For a sample product, 16 out of 30 features are selected: *T*10, *T*20, *MA*3, *MA*5, *MA*15, *RDP*3, *BIAS*10, *BIAS*15, *ROC*10, *Disparity*5, *Disparity*10, *OSCP*5, *IML*, *IMM*, *IMS*1, *IMS*2.

Table 3. Evaluation of the proposed intermittency features.

Performance criteria		**SVR without IF** **(only Table 1)**	**SVR with IF** **(Tables 1 and 2)**
RMSE	Maximum	4.355	4.121
	Minimum	0.000	0.000
	Average	0.911	0.709
	Standard deviation	0.728	0.622
MSE	Maximum	18.969	16.979
	Minimum	0.000	0.000
	Average	1.361	0.889
	Standard deviation	2.689	2.070
MAD	Maximum	2.656	1.923
	Minimum	0.000	0.000
	Average	0.541	0.367
	Standard deviation	0.493	0.376

4.3.3. Forecasting results

The performance of the proposed clustering-based forecasting (C-MARS+SVR) methodology is compared with other forecasting models: S-SVR, S-MARS+SVR, and C-SVR. In S-SVR and S-MARS+SVR, a separate forecasting model is developed for each product, so 98 forecasting models are generated. In C-SVR and C-MARS+SVR, a forecasting model is developed for each cluster, so the number of models is equal to the number of clusters. Also, cluster representatives are calculated using medoid and DBA methods.

Table 4 summarizes the numerical results. For all forecasting error measures, the best results are observed in S-MARS+SVR. Except MSE measure, S-SVR follows S-MARS+SVR. On the other hand, 98 models need to be trained and tested for both S-SVR and S-MARS+SVR. As a remedy, clustering is used to reduce the number of models. Among three clustering results, C-MARS+SVR with *k*=27 and DBA provides the minimum error. The representative selection method and the number of clusters affect the forecasting errors. In another word, when *k*=27, DBA yields smaller forecasting errors compared to medoid approach. For other *k* values, medoid approach provides smaller error values. As the number of clusters decreases, all error measures increase. Therefore, the sales patterns of the products are better identified for larger number of clusters.

The use of MARS improves the forecasting errors for individual models and for k = 27. It eliminates the redundant variables and results in a less complex model.

Table 4. Comparison of the forecasting models.

Method	**Complexity**	**Average**		
		RMSE	**MSE**	**MAD**
S-SVR	2817	0.942	3.529	0.502
S-MARS+SVR	1072	**0.885**	**2.811**	**0.474**
C-SVR with *k*=7 and medoid	195	1.375	4.471	0.916
C-SVR with *k*=7 and DBA	207	1.588	5.315	1.037
C-MARS+SVR with *k*=7 and medoid	80	1.376	3.905	0.916
C-MARS+SVR with *k*=7 and DBA	**69**	1.622	4.700	1.091
C-SVR with *k*=16 and medoid	464	1.262	4.088	0.820
C-SVR with *k*=16 and DBA	477	1.354	4.355	0.816
C-MARS+SVR with *k*=16 and medoid	143	1.239	3.440	0.809
C-MARS+SVR with *k*=16 and DBA	132	1.390	3.869	0.885
C-SVR with *k*=27 and medoid	781	1.248	3.991	0.791
C-SVR with *k*=27 and DBA	781	1.220	3.957	0.746
C-MARS+SVR with *k*=27 and medoid	254	1.185	3.248	0.754
C-MARS+SVR with *k*=27 and DBA	241	1.165	3.236	0.724

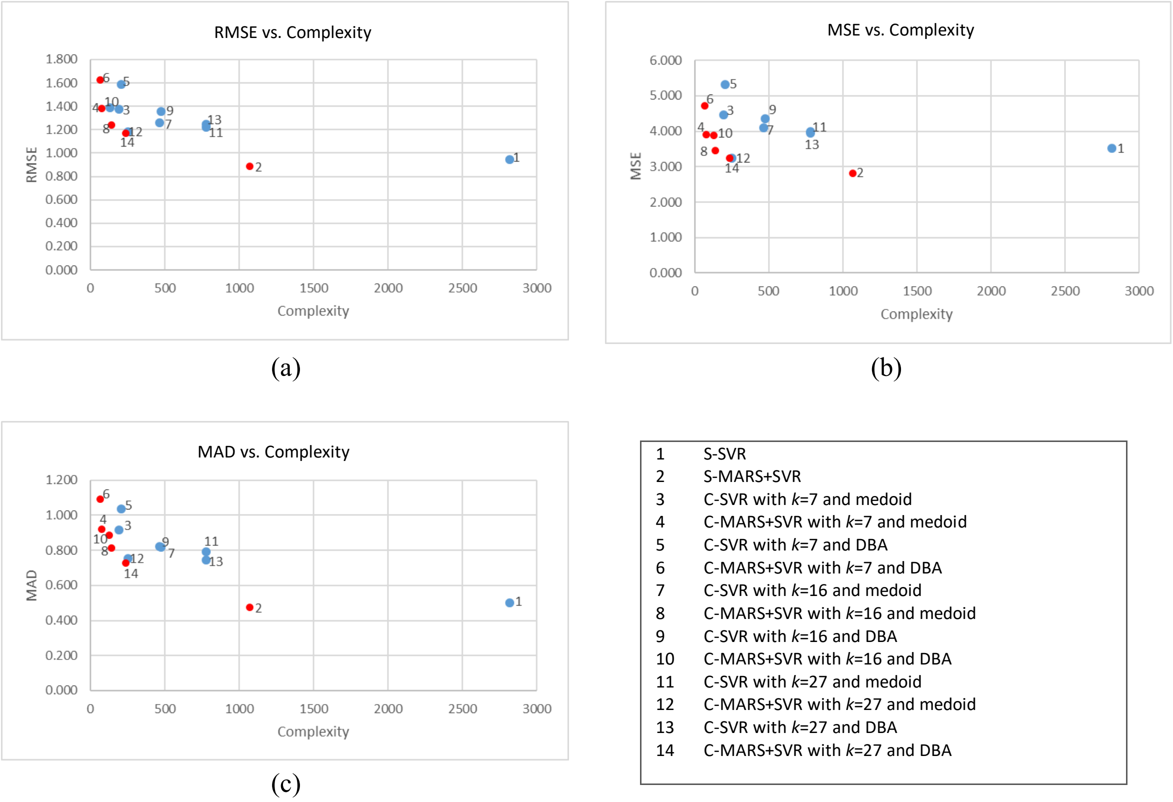
4.4. Discussion

In this paper, the aim is to obtain an accurate forecasting model with low complexity. Hence, two criteria are used to evaluate the forecasting models. The first criterion is the forecasting error, i.e. RMSE, MSE, or MAD. The second criterion is the complexity, i.e. the number of forecasting models and the number of features used in the model. Hence, the best forecasting model is selected within the multiple criteria (attribute) decision making framework. In this context, a solution is *dominated* if there are other solutions that are better than it in at least one criterion and as good as it in other criteria (Yoon and Hwang, 1995). If a solution is not dominated by other solutions, it is called *non-dominated solution* (Yoon and Hwang, 1995).

The non-dominated solutions obtained with respect to the two criteria are presented in Fig. 7. Based on average RMSE and complexity, the non-dominated solutions are S-MARS+SVR, C-MARS+SVR with *k*=7 and medoid, C-MARS+SVR with *k*=7 and DBA, C-MARS+SVR with *k*=16 and medoid, and C-MARS+SVR with *k*=27 and DBA. Based on average MSE and complexity, S-MARS+SVR, C-MARS+SVR with *k*=7 and medoid, C-MARS+SVR with *k*=7 and DBA, C-MARS+SVR with *k*=16 and medoid, C-MARS+SVR with *k*=16 and DBA, and C-MARS+SVR with *k*=27 and DBA are the non-dominated solutions. The non-dominated solutions for average MAD and complexity are the same ones obtained for average MSE and complexity.

On one hand, S-MARS+SVR provides the minimum errors with a high complexity. On the other hand, C-MARS+SVR with *k*=27 and DBA provides reasonable errors with less complexity. That is, 77.5% of reduction in the complexity results in 15.1% of increase in MSE. Meanwhile, the RMSE and MAD values increase by 31.6% and 52.7%, respectively.

To sum up, the decision maker can select the proper forecasting method considering the balance of error and complexity. Also, the proposed approach can be applicable for the newly released products. The sales of the product can be forecasted using the cluster representative.



(a)

(b)

(c)

1	S-SVR
2	S-MARS+SVR
3	C-SVR with *k*=7 and medoid
4	C-MARS+SVR with *k*=7 and medoid
5	C-SVR with *k*=7 and DBA
6	C-MARS+SVR with *k*=7 and DBA
7	C-SVR with *k*=16 and medoid
8	C-MARS+SVR with *k*=16 and medoid
9	C-SVR with *k*=16 and DBA
10	C-MARS+SVR with *k*=16 and DBA
11	C-SVR with *k*=27 and medoid
12	C-MARS+SVR with *k*=27 and medoid
13	C-SVR with *k*=27 and DBA
14	C-MARS+SVR with *k*=27 and DBA

Fig. 7. Non-dominated solutions with respect to the forecasting errors and complexity.

5. CONCLUSION

This study presents a forecasting methodology for the companies that offer high product variety. The proposed methodology particularly focuses on sales data with unequal lengths and intermittency. For this purpose, the methodology combines various data mining tasks, i.e. clustering, feature extraction, feature selection, and prediction.

First, clustering is performed to determine the products having similar sales patterns. In this context, DTW is adopted to determine the dissimilarities for the products having different release and phase-out times. Second, in addition to the features that consider trend, volatility, and so on, new features are proposed to characterize intermittency. Additionally, for each cluster representative, features are selected with MARS. Finally, SVR is used for forecasting sales.

The proposed methodology is implemented in a forklift distributor company. Numerical results indicate that the proposed approach provides a reasonable level of accuracy with low complexity. Meanwhile, the non-dominated solutions with respect to the forecasting error and complexity are identified, and trade-offs between the two criteria are presented. The decision maker can select an appropriate model based on his/her preferences.

The proposed approach can be applied to a wide variety of companies such as retailers of fast fashion. Further studies can focus on the incremental updates of the clustering and prediction results. Additionally, new dissimilarity measures can be developed for intermittent data.

REFERENCES

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. The Fourth International Conference Foundations of Data Organization and Algorithms, 69-84. doi:10.1007/3-540-57301-1_5

Bala, P.K. (2012). Improving inventory performance with clustering-based demand forecasts. Journal of Modelling in Management, 7(1), 23-37. doi:10.1108/17465661211208794

Bao, Y., Wang, W., & Zhang, J. (2004). Forecasting intermittent demand by SVMs regression. IEEE International Conference on Systems, Man and Cybernetics, 1, 461–466. doi:10.1109/icsmc.2004.1398341

Bao, Y., Wang, W., & Zou, H. (2005). SVR-based method forecasting intermittent demand for service parts inventories. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 604–613. doi:10.1007/11548706_64

Berndt, D., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. Workshop on Knowledge Knowledge Discovery in Databases, 398, 359–370.

Biscarri, F., Monedero, I., García, A., Guerrero, J. I., & León, C. (2017). Electricity clustering framework for automatic classification of customer loads. Expert Systems with Applications, 86, 54–63. doi:10.1016/j.eswa.2017.05.049

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Brown, R. G. (1959). Statistical forecasting for inventory control. McGraw/Hill.

Chen, I. F., & Lu, C. J. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing. Neural Computing and Applications, 28(9), 2633–2647. doi:10.1007/s00521-016-2215-x

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. Operational Research Quarterly, 23(3), 289-303. doi:10.2307/3007885

Dai, W., Chuang, Y.-Y., & Lu, C.-J. (2015). A Clustering-based sales forecasting scheme using support vector regression for computer server. Procedia Manufacturing, 2, 82–86. doi:10.1016/j.promfg.2015.07.014

Das, S., & Padhy, S. (2012). Support vector machines for prediction of futures prices in Indian stock market. International Journal of Computer Applications, 41(3), 22–26. doi:10.5120/5522-7555

Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19(1), 1–67.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques. San Francisco, CA, Morgan Kaufmann.

Hautamaki, V., Nykanen, P., & Franti, P. (2008). Time-series clustering by approximate prototypes. 19th International Conference on Pattern Recognition, 1–4. doi:10.1109/ICPR.2008.4761105

Hua, Z., & Zhang, B. (2006). A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. Applied Mathematics and Computation, 181(2), 1035–1048. doi:10.1016/j.amc.2006.01.064

Huber, J., Gossmann, A., & Stuckenschmidt, H. (2017). Cluster-based hierarchical demand forecasting for perishable goods. Expert Systems with Applications, 76, 140–151. doi:10.1016/j.eswa.2017.01.022

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8), 651–666. doi:10.1016/j.patrec.2009.09.011

Keogh, E. (1997). A fast and robust method for pattern matching in time series databases. In: Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, pp. 578–584.

Kumar, V., & Rathi, N. (2011). Knowledge discovery from database using an integration of clustering and classification. International Journal of Advanced Computer Science and Applications, 2(3), 29–33. doi:10.14569/ijacsa.2011.020306

Kuo, R. J., & Li, P. S. (2016). Taiwanese export trade forecasting using firefly algorithm based K-means algorithm and SVR with wavelet transform. Computers and Industrial Engineering, 99, 153–161. doi:10.1016/j.cie.2016.07.012

Levis, A. A., & Papageorgiou, L. G. (2005). Customer demand forecasting via support vector regression analysis. Chemical Engineering Research and Design, 83(8 A), 1009–1018. doi:10.1205/cherd.04246

Lu, C.-J. (2014). Sales forecasting of computer products based on variable selection scheme and support vector regression. Neurocomputing, 128, 491–499. doi:10.1016/j.neucom.2013.08.012

Lu, C. J., and Kao, L. J. (2016). A clustering-based sales forecasting scheme by using extreme learning machine and ensembling linkage methods with applications to computer server. Engineering Applications of Artificial Intelligence, 55: 231–238. doi:10.1016/j.engappai.2016.06.015

Lu, C. J., Lee, T. S., and Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. Decision Support Systems, 47(2): 115–125. doi:10.1016/j.dss.2009.02.001

Lu, C. J., Lee, T. S., & Lian, C. M. (2012). Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. Decision Support Systems, 54(1), 584–596. doi:10.1016/j.dss.2012.08.006

Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. Computers & Industrial Engineering, 109, 233–252. doi:10.1016/j.cie.2017.04.017

Nalbantov, G., Groenen, P. J., & Bioch, J. C. (2005). Support vector regression basics. Medium Econometrische Toepassingen, 13(1), 16-19.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition, 44(3), 678–693. doi:10.1016/j.patcog.2010.09.013

Thissen, U., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). Using support vector machines for time series prediction. Chemometrics and Intelligent Laboratory Systems, 69(1–2), 35–49. doi:10.1016/S0169-7439(03)00111-4

Thomassey, S., & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems, 42(1), 408–421. doi:10.1016/j.dss.2005.01.008

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer Verlag. doi:10.1007/978-1-4757-2440-0

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. Management Science, 6(3), 324–342. doi:10.1287/mnsc.6.3.324

Wisner, J. D., Tan, K. C., & Leong, G. K. (2015). Principles of supply chain management: A balanced approach. Cengage Learning.

Yoon, K. P., & Hwang, C. L. (1995). Multiple attribute decision making: an introduction. Sage publications.

Yu, X., Qi, Z., & Zhao, Y. (2013). Support vector regression for newspaper/magazine sales forecasting. Procedia Computer Science, 17, 1055–1062. doi:10.1016/j.procs.2013.05.134

Zuo, Y., Ali, A. B. M. S., & Yada, K. (2014). Consumer purchasing behavior extraction using statistical learning theory. Procedia Computer Science, 35, 1464–1473. doi:10.1016/j.procs.2014.08.209