K-means clustering combined with principal component analysis for material profiling in automotive supply chains

Abstract. At a time where available data is rapidly increasing in both volume and variety, descriptive Data Mining (DM) can be an important tool to support meaningful decision-making processes in dynamic Supply Chain (SC) contexts. Up until now, however, scarce attention has been given to the application of DM techniques in the field of inventory management. Here, we take advantage of descriptive DM to detect and grasp important patterns among several features that coexist in a real-world automotive electronics SC. Concretely, Principal Component Analysis (PCA) is employed to analyze and understand the interrelations between ten quantitative and dependent variables in a multi-item/multi-supplier environment. Afterwards, the principal component scores are characterized via a K-means clustering, allowing us to classify the samples into four clusters and to derive different profiles for the multiple inventory items. This work provides evidence that descriptive DM contributes to find interesting feature-patterns, resulting in the identification of important risk profiles that may effectively leverage inventory management for superior performance.

Keywords: Supply chain · Data mining · K-means clustering · Principal component analysis (PCA).

1 Introduction

Aiming to cope with the fast and real time changes on the modern business environments, it is fundamentally important to perceive Supply Chain (SC) dynamics [1], especially at a time where there is a pressing need for SC integration [2]. Bearing in mind that organizations are commonly structured in SCs [3], Supply Chain Management (SCM) plays a paramount role in promoting their success, achieving their objectives and, above all, guaranteeing customer satisfaction [4]. In this context, the inventory management process is considered to be an important driver for the success of a company, notwithstanding the challenges related to demand and supply uncertainty attached thereto [5]. In the literature, this process is closely bound up with the volatility of inventory components, namely raw materials, work-in-process, and finished goods [6].

In highly volatile and dynamic markets, as in the case of the automotive sector, SC managers tend to order components well beforehand in order to avoid stock-outs. As a corollary, this leads to excess inventory, as well as to increased holding costs and higher risks of product obsolescence. Therefore, it is essential to strike the proper balance between stock-outs and excess inventory [6], so that the customer service level is maintained whilst minimizing total SC costs. Thus, for a given inventory component, a comprehensive knowledge of its typical profile, based on the dynamic interplay between the various parameters associated with it, might provide important insights on how to manage it. Moreover, due to the fact that raw material inventory is directly influenced by interactions with suppliers (see [6] and references cited therein), the buyer-supplier relationship can also be enhanced during this profiling process. Indeed, special importance should be attached to this mutual relation. Following the reasoning of Talluri and Sarkis [7], buyers should monitor supplier performance in such a way that the information derived from this monitoring process can be shared with suppliers, in order to encourage them to take actions able to meet the requests of the buyers.

Nonetheless, although the dynamic behavior is an inherent feature within any SC, especially regarding the stochasticity of SC parameters, it tends to be undervalued or even neglected, particularly with regard to risk assessment [8]. This, together with the complex business environments characterized by the rapid growth of generated data [9], puts pressure on companies to take advantage of new approaches and techniques able to support decision-making processes. At this point, the ultimate purpose relates to the extraction of valuable insights from raw data, in order to generate new competitive advantages. The application of these techniques is particular interesting in the framework of the automotive electronics sector, for which estimates point to a 8% growth forecast over 2017-2024 with an associated market share of more than $390 billion by 2024 [10].

Increasingly, Data Mining (DM) techniques have been proposed to improve SC processes, for instance relating to the ranking, selection and evaluation of suppliers (see, e.g., [11–14]). Up until now, however, the application of DM techniques in the field of inventory management have not been fully explored, as other aspects remain to be investigated. Indeed, this statement was recently emphasized by Moharana and Sarmah [15]. For example, Tsai et al. [16] introduces an association clustering algorithm capable to group a large number of products with identical demands in a hierarchical fashion, under the can-order policy model. Simulation experiments showed the benefits of the proposed approach when compared with different replenishment models in terms of total profit, sales revenue, as well as holding, shortage and ordering costs. By contrast, Aqlan [17] applied K-means clustering to group inventory parts according to different features. The obtained clusters served as a guideline for warehouse space optimization. Kartal et al. [18] consider the joint application of multi-criteria decision making approaches with machine learning algorithms in the field of multi-attribute inventory classification (MCIC). The proposed approach was conducted in a real-world automotive production company in Turkey and revealed to be applicable to multiple inventory structures. The benefits resultant of the application of supervised machine learning methods for MCIC purposes are also highlighted in the research conducted by Lolli et al. [19].

Focusing on methods that do not require a-priori knowledge of underlying patterns, also known as unsupervised methods [20], this paper addresses the problem of identifying different profiles for multiple inventory components based on the interplay between several variables collected from a real-world automotive SC with multiple suppliers. For that, descriptive Data Mining (DM) techniques [21] are employed. Concretely, the mathematical relationship between ten quantitative and dependent variables is firstly studied by taking advantage of Principal Component Analysis (PCA). Afterwards, K-means clustering based on the principal component (PC) scores is used to identify and characterize different inventory component profiles. The derived clusters are validated via 10-fold cross-validation using different benchmark clustering models and validity indexes, stressing the relevance of this work in bridging the literature gap related to the application of DM approaches in the field inventory management, already pointed by [15]. By simplifying the complexity in the dataset without much loss of information, this work contributes to extant literature by proposing a descriptive DM approach that acts as a monitoring mechanism for the status of multiple inventory component groups in real-world SC contexts. Moreover, it can be used by SC managers and practitioners as a supporting tool for the decision-making process, whilst contributing to the continuous improvement of inventory management.

The rest of the paper is organized as follows. Section 2 presents the real-world collected data, as well as the selected unsupervised learning models. In Section 3 we describe the PCA framework. Next, the numerical results derived from the application of K-means based on PCA are analyzed and discussed in Section 4. Finally, conclusions are carried out in Section 5.

2 Materials and methods

2.1 Dataset

A total of 9806 records, associated with 59 inventory components and 39 worldwide suppliers, were collected from a major automotive electronics supply chain, located in Europe, for the years of 2016 and 2017. For reasons of confidentiality, we have omitted the company name. Each record represents information of a given component for a particular day and supplier. Concretely, 12 features were measured, from which 10 of them are quantitative and dependent variables. After data cleansing, the company managers manually grouped each component in one of 6 different categories, namely: “high runner” (4818 records), for fast-moving components; “special freights” (1202 records), referring to products with high marginal propensity to incur in a special freight (e.g., due to stock-outs events); “critical” (1324 records), to represent problematic components (e.g., in terms of quality issues or highly demand fluctuation); “stable” (934 records), to identify components without deviant behaviors; “commodity”, to represent undifferentiated components (577 records), and “common among plants” to represent components that are used in several company plants (951 records). For this particular dataset, we found that the categories are non-overlapping, i.e, each component belongs to one and only one category. However, it should be noted that further datasets can contain components belonging to more than one category. A short descriptive analysis of each feature is provided in Table 1.

Table 1: Basic descriptive analysis of the dataset.

Feature	Notation	Domain	Mean/Mode	SD	Description
qty.rec.	F_1	$[1, 134136]$	3572.08	7405.94	Stock quantity received
saf.time	F_2	$[1, 15]$	3.26	2.16	Time buffer added to the supply lead time that pushes a delivery order earlier
val.stock	F_3	$[0, 791907.2]$	30477.99	65237.26	Monetary value of stock on-hand
cons.stock	F_4	$[0, 14178]$	1320.37	1670.86	Quantity of stock expected to be consumed
supp.otd	F_5	$[1, 100]$	75.85	25.62	Supplier On-Time Delivery (OTD) score
wh.occup	F_6	$[0, 82]$	10.71	10.78	Number of warehouse bins occupied
stock	F_7	$[0, 861172]$	16300.98	53068.48	Quantity of stock on-hand
moq	F_8	$[0, 72000]$	1285.86	6294.94	Agreed Minimum Order Quantity (MOQ) with supplier
supp.lt	F_9	$[1, 38]$	7.34	9.01	Time interval between ordering and receiving an item order
nr.end	F_{10}	$[1, 109]$	21.63	28.28	Number of end-items that make use of the component in their Bill of Materials
rm.cat.	F_{11}	-	High runner	-	Material category ({"Stable", "High runner", "Special freights", "Critical", "Commodity", "Common among plants"})
geo.loc.	F_{12}	-	Portugal	-	Geographical location of the supplier (e.g., {"Germany", "Spain", "Portugal", "China", "Japan"})

SD = Standard Deviation.

2.2 Selected unsupervised learning models

Two unsupervised learning methods, namely Principal Component Analysis (PCA) and K-means clustering, were tested in order to describe, in a quantitative fashion, the relationships between the variables $F_i, i = 1, \ldots, 10$, in the data matrix $\mathbf{X}_{9806 \times 10}$. A short theoretical introduction of both methods is provided as follows (see [22, 23] and references cited therein for more detailed information regarding these topics).

Principal Component Analysis (PCA). As a descriptive and multivariate statistical technique, PCA was firstly studied by [24] and [25]. In short, PCA intends to compress the dimension of a given dataset, whilst minimizing statistical information loss [26]. Let $\mathbf{X}$ be a $(n \times p)$ data matrix, with n observations and p features. Considering $\mathbf{X}$ as a p-dimensional vector of random features, we denote the covariance matrix of $\mathbf{X}$ by $\sum_{\mathbf{X}}$, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and eigenvectors $\alpha_1, \alpha_2, \ldots, \alpha_p$. PCA seeks to determine a new set of q variables $\mathrm{PC}_i, i = 1, \ldots, p$ where $(q \ll p)$, called *principal components*, that represent linear combinations of the original features and are uncorrelated with each other in a descending order of relevance in terms of total variance explained [27]. Each component can be then interpreted according to the inter-correlated variables that comprise it. Moreover, α_{ij} is called the PC *loading* and represents the relative contribution of the jth original feature for the ith PC. The elements of the linear combinations PC_i are commonly referred as PC *scores*. It is noteworthy that the diagonal elements of $\sum_{\mathbf{X}}$, $\lambda_i, i = 1, \ldots, p$, traduce the variance explained by each PC and decrease monotonically from PC_1 to PC_p. In this regard, a natural problem that may arise relates to determine how many PCs should be retained. Albeit this problem continues to be unresolved, some methods have been proposed in the literature to tackle it [27]. Typically, two approaches are often used to select the number of PCs to retain. The first one consists in selecting the PCs in which $\lambda_i > 1, i = 1, \ldots, p$. The second involves retaining the largest number of PCs that, together, account from 70% to 90% of total variance explained in the dataset. Nonetheless, this interval may vary depending on the data concerned [22].

***K*-means clustering.** Driven by the studies of [28–31], K-means is an iterative descent clustering method [32], considered to be the most widely used algorithm for partitional clustering [33]. Let $X = \{x_{ij}\}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$, be the set of observations in the data matrix $\mathbf{X}$ to be assigned into a K-dimensional set $C = \{C_k, k = 1 \ldots, K\}$. Given the a-priori number of desired clusters K, the main idea of K-means is to partition the n p-dimensional observations into K clusters in such a way that the total within-cluster variation, $W(C_k)$, is minimized. Following the formulations of [34], the within-cluster variation for the kth cluster is typically expressed as the sum of all the pairwise squared Euclidean distances between the observations in the kth cluster, divided by the total number of observations, $|C_k|$, therein contained. This reasoning can be translated into the following optimization problem

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} . \tag{1}$$

Despite of this optimization problem be NP-hard, as pointed in [23], a local optimum can be derived by taking advantage of a simple algorithm in which each observation is assigned to the cluster whose centroid, defined by $\sum_{i \in C_k} x_i |C_k|^{-1}$, is closest (in our case in terms of the Euclidean metric). The computation of the K-means depends on three pre-specified parameters, namely: (1) the number of clusters, K, for which there is no theoretical approach to define it [23]; (2) the distance metric considered - typically the Euclidean distance metric, notwithstanding other distance metrics can be used (e.g., Mahalanobis and Gower); and (3) the initial cluster assignment (also called cluster initialization). Here, it is a common practice to test different random initial assignments for a predefined value of K, inasmuch as K-means does not provide a global optimum. At the end, it is chosen the solution for which the optimization problem (1) is minimized [23, 34].

In this study, the optimal number K is selected via the R-squared (RS) [35] and the prediction strength [36] validity indexes. Algebraically, the RS index is defined as $RS = 1 - SS_w/SS_t$, where SS_w and SS_t are the sum of squares within each group and the total sum of squares for the whole dataset, respectively. The RS index takes values in the compact interval $[0, 1]$. At this point, if the value of RS is 0 there exists no significant differences between clusters. By contrast, values of RS close to 1 indicate a well separation between clusters, as well as a high degree of homogeneity intra-cluster. Regarding the prediction strength approach, it treats clustering as a supervised classification problem in which the main idea is to cluster both train and test data into K clusters and compute, for each test cluster, the proportion of observation pairs therein contained that are also classified into the same cluster by the training centroids (see [36] for details).

3 Modeling framework

The numerical experiments presented throughout this section were conducted in the **R** programming language [37] with suitably selected packages.

Firstly, we adopt PCA in order to transform a set of correlated variables into a smaller set of linearly uncorrelated variables, known as PCs, which retain the most relevant information from the original dataset whilst minimizing information lost. Therefore, with the application of PCA we intend to identify the most relevant logistic information patterns from a dimensional feature subspace with less than the number of original features. In the literature, several applications of PCA have been proposed in the context of SCM, showing relevant benefits on the supplier selection problem in multi-item/multi-supplier environments [38] or on the extraction of the most relevant sustainability indicators to conduct eco-efficiency performance analyses in industrial companies [39]. PCA can also contribute to the identification of operational risk sources (see, e.g., [40]) and, for our case in particular, to better comprehend the risk profiles of the different inventory items according to the logistic features associated with them. With this knowledge base, we expect that company experts can develop more effective action plans to improve and support the inventory control decision-making process. Secondly, the PC scores are used as input features for K-means clustering. Following this approach it is intended to apply K-means clustering on a low-dimensional dataset rather than on the original 10-dimensional feature subspace. This represents a relevant advantage in real-world business contexts inasmuch as it facilitates the use of this approach by improving its interpretability. Indeed, it is common to combine these two unsupervised strategies for data dimensional reduction purposes (see [41–43] and references cited therein).

In the next subsections, we provide the details of both PCA and K-means experimental setups.

3.1 PCA experimental setup

The features $F_i, i = 1, \ldots, 10$, presented in Table 1 have different units of measurement. At this point, the use the covariance matrix in the original data space would give greater weight to features with more variance and, in contrast, less weight to features with smaller variance. Therefore, since we intend to treat all input features on an equal basis, we preferred the use of the correlation matrix rather than the covariance matrix [22], since with the former all features are typically standardized to unit variance. Note that the correlation matrix of the original data boils down to the covariance matrix of the standardized data. In this context, performing PCA on the standardized data is commonly referred to correlation matrix PCA [26]. Moreover, since classical PCA is not robust to outliers and noise data, we considered a Minimum Covariance Determinant (MCD)-based PCA (see [44]). The MCD method adopts a highly robust estimator of multivariate locator and scatter and has been explored to develop robust multivariate approaches (the reader is referred to [45] for details). Following this approach, it is expected that the results derived by PCA based on a robust correlation matrix are not overly influenced by the presence of pre-existing outliers [44]. Concerning the selection of the number of PC included, there exists a trade-off between increasing variance explained while reducing the number of PCs containing irrelevant information or noise. Following common yet subjective practice [26], we retain the components which account for at least 70% cumulative explained variance. This leads to the selection of the first 4 PCs, accounting for approximately 78% of cumulative total variance explained in the dataset (see Fig. 1).

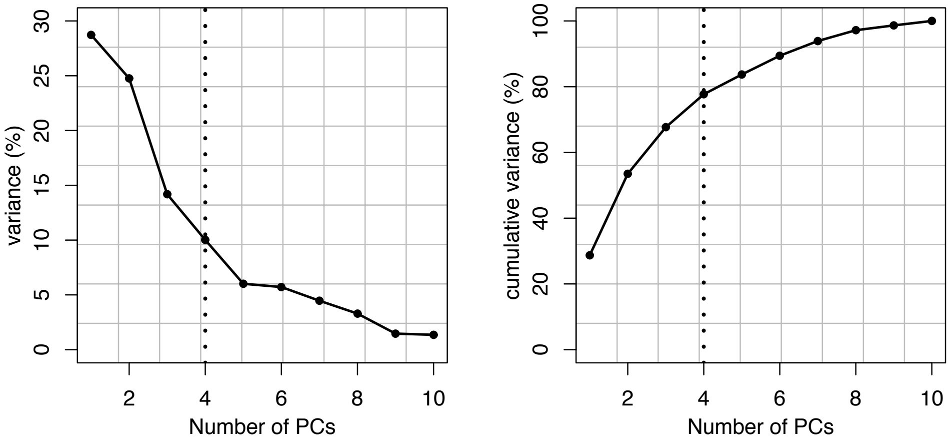

Fig. 1: Variance (left) and cumulative variance (right) explained as a function of the number of PCs.

Under the Kaiser's rule [46], the exclusion of the remaining PCs is justified by the fact that the respective eigenvalues are not equal to or greater than one. In this regard, we found through background analyses (not presented) that the different samples showed a strong overlap on the higher-order principal components, which represent the remaining 22% of the variability. Thus, this indicates that there is no relevant logistic information contained therein, and the inclusion of such higher-order principal components would essentially represent noise.

On the other hand, since PCs are linear combinations of all the dataset features, we identified, for each PC, which features can be discarded for the sake of interpretability while preserving as much as possible statistical information. Typically, this identification is merely based on the magnitude of the feature loadings, neglecting those with low magnitude, which can be potentially misleading (see [47]). Thus, in a bid to reduce the subjective nature inherent to the interpretation of PCs [47], we also analyzed the relationships between the features F_i , $i = 1, 2, \ldots, 10$, and the different PCs via *correlation circles*, in which the features are represented as points in the PC space using their correlations with each PC as coordinates [27]. Figure 2 plots the correlation circles for the first four PC dimensions.

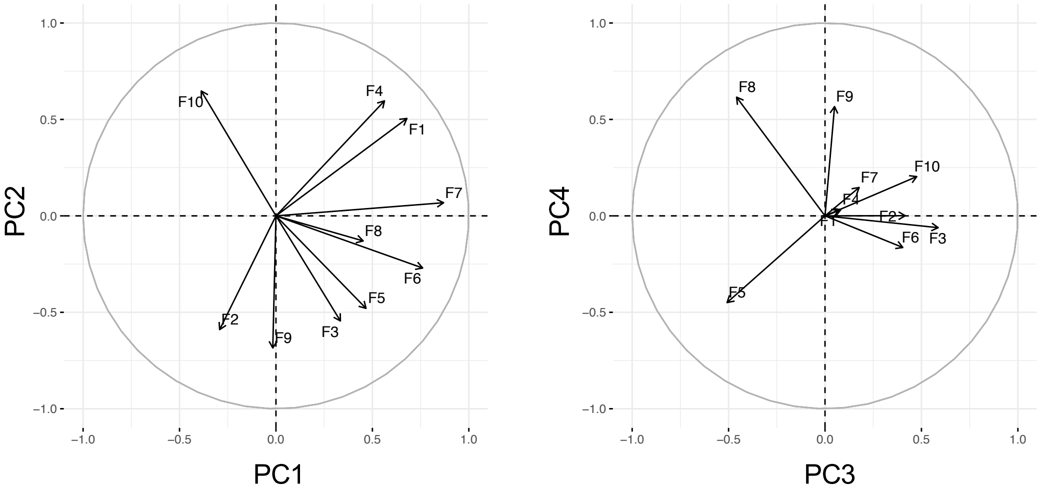

Fig. 2: Correlation circles for the first and second (left), and third and fourth (right) PCs.

In both circles, particular attention should be given to the distance between the features and the origin. The closer a feature is to the unit circle, the higher its relevance for interpreting the concerned components in detriment of ones closer to the origin. In addition, two arbitrary features projected in the PC space are said to be positive (negative) correlated variables if they are pointing in the same (opposite) direction. In contrast, they are said to be unrelated if they are orthogonal to each other. By way of example, examination of the left circle plotted in Fig. 2 shows that the first PC (PC_1) reflects the components' inventory levels since it is mostly correlated with the stock quantity received (F_1), the warehouse occupation (F_6) and the quantity of stock on-hand (F_7). Yet, with the exception of the stock quantity received, these features seem to have no strong correlation with PC_2, which essentially contrasts the safety time (F_2), the supplier lead time (F_9) and the monetary value of stock with both the stock quantity received (F_1) by the organization and the number of end-items that make use of that specific material in their Bill of Materials (F_{10}). Combined, the results derived from the correlation circles together with both the magnitude and signs of the PC loadings allow us to derive truncated PCs ($\mathrm{PC}_i^{tr}, i = 1, \ldots, 4$). Each PC_i^{tr} is described as follows:

$$\mathrm{PC}_1^{tr} = 0.4001F_1 + 0.3326F_4 + 0.4476F_6 + 0.5120F_7 + 0.2661F_8 \tag{2}$$

$$\mathrm{PC}_2^{tr} = 0.3208F_1 - 0.3751F_2 - 0.3480F_3 - 0.4363F_9 \tag{3}$$

$$\mathrm{PC}_3^{tr} = 0.4916F_3 - 0.3853F_8 \tag{4}$$

$$\mathrm{PC}_4^{tr} = 0.6159F_8 + 0.5666F_9 - 0.0634F_3 - 0.4513F_5 \tag{5}$$

The selected subsets of features and the interpretation of each PC_i^{tr} appear summarized in Table 2. Note that the interpretation of each PC_i^{tr} depends both on the magnitude and signs of the variable loadings. For instance, the algebraically formulation of PC_1^{tr} given by Eq. (1) shows that all the variables that comprise it are inventory-related and the corresponding loadings are positive. Thus, this suggests that PC_1^{tr} can be interpreted as a weighted average of the inventory level, where samples with high PC_1^{tr} scores exhibit high inventory levels, and vice versa. In contrast, PC_2^{tr} comprises three variables with negative loadings and one variable with positive loading. Therefore, high values of PC_2^{tr} reflect the contrast of the stock quantity received with the safety time, supply lead time and stock monetary value of the six different component categories previously defined.

Overall, it is noteworthy that the truncated PCs are easier to be interpreted when compared to the original PCs, due to the smaller subset of features which constitute them. Moreover, whatever the truncated PC concerned, its correlation with the original PC is quite reasonable ($\geq$ 0.8759), which corroborates the quality of approximation of the four extracted PCs using the truncated components.

Table 2: Summary of the truncated PCs.

PC_i^{tr}	Subset of features	$\mathrm{Corr}(\mathrm{PC}_i^{tr}, \mathrm{PC}_i)$	Interpretation
$i = 1$	$\{F_1, F_4, F_6, F_7, F_8\}$	0.8914	Weighted average of F_1, F_4, F_6, F_7, F_8
$i = 2$	$\{F_1, F_2, F_3, F_9\}$	0.9401	Contrast between F_1 and F_2, F_3, F_9
$i = 3$	$\{F_3, F_8\}$	0.8759	Contrast between F_3 and F_8
$i = 4$	$\{F_3, F_5, F_8, F_9\}$	0.8833	Contrast between F_3, F_5 and F_8, F_9

3.2 K-means experimental setup

When choosing the initial centroids and selecting the number of clusters K to retain, multiple random initial configurations are typically tested. In fact, this approach is considered to be the most widely used [48]. However, apart from this strategy, there exist other initialization methods suitable for this purpose (see, e.g., [48]). In this work, 24 sets of cluster centers were obtained via the Ward's hierarchical agglomerative clustering method [49]. Then, the derived centroids are used as starting centroids in the regular K-means approach. Former studies had already pointed the benefits of this adoption for obtaining good clusters [48, 50]. In this process we considered the Euclidean metric and the Ward2 algorithm [51] implemented in the **hclust** function within the **R** package **stats**. Based on the RS indexes resulting from the different initializations, the number of clusters was then set at $K = 4$ (left of Fig. 3). This choice is corroborated by the average prediction strength value (right of Fig. 3) attained for $K = 4$ ($\overline{ps}|_{K=4} = 0.8524$ with cutoff $= 0.8$ and 100 resampled datasets), which represents a proper threshold for obtaining well separated clusters [36].

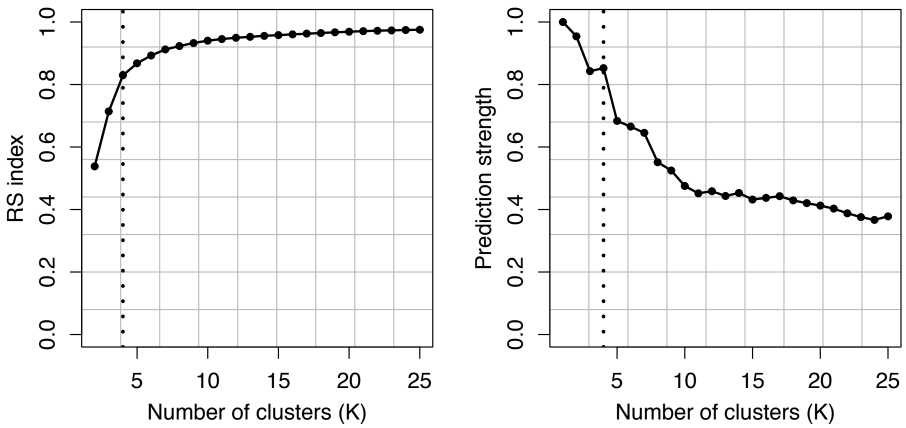

Fig. 3: R-squared and prediction strength indexes as a function of K.

4 Results

We hypothesized that PCA can provide valuable information to identify relevant relationships among samples, and to collect some information regarding the logistic behavior of the various components over time. Thus, we studied the changes of PC scores in the first two PC dimensions, which explain roughly 54% of the variability, with increasing number of samples from the first semester of 2016 (S1) until the end of 2017 (S4) (Fig. 4). To confirm this hypothesis, four time frames are considered in subsequent analyses: S1, containing numerical data related to the first semester of 2016; [S1, S2], representing samples related to the entire year 2016; [S1, S3], referring to collected data from S1 to the first semester of 2017; and [S1, S4], containing the whole dataset from 2016 to 2017.

In Fig. 4, each sample is related to a particular component category. In its turn, since the categories are non-overlapping, each category is coded with a specific color. Figure 4 shows that the samples distribution on the PC subspaces differs over the time frames considered. In particular, it reveals that with an increase of the number of samples from [S1, S3] to [S1, S4] some commodities are no longer located on the positive semi-axis of PC_2, meaning that the average stock quantity received related to those components decreased substantially in that period. In addition, over the year 2016 ([S1, S2]), components prone to special freights were mainly located on the negative semi-axis of PC_2, attaining minimum PC scores of close to -20. Yet, by gathering the data of 2016 together with the first semester of 2017 ([S1, S3]) we have noticed that the same PC scores have become increasingly negative over the PC_2, which have translated into higher safety times, supply lead times and monetary stock values for some observations that fall within that specific component category. Overall, by projecting the PC scores onto the first two PCs over the time window considered, it was possible to detect important feature-patterns that provide company managers with valuable insights regarding the status of the different types of component in that period. Nevertheless, the different samples showed a strong overlap on these two PC-dimensions, making it difficult to identify any further relevant information among the projected samples. In this case, PCA fails to properly separate the samples in such a way so as to be able to extract further insights from the dataset. Yet, we were interested in investigate if this apparent drawback of PCA is motivated by an incorrect classification of the samples during the data preparation stage, in the sense that there may be samples from distinct component categories that, due to their similarity, could be grouped into the same category or cluster.

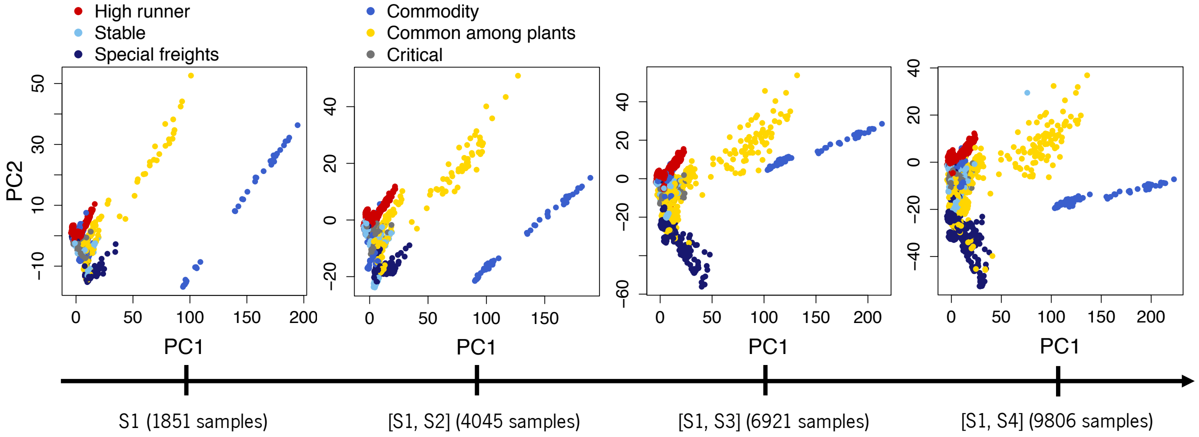

Fig. 4: Evolution of the first two PC scores with increasing number of samples over four distinct time frames.

4.1 Visualizing PC scores via *K*-means

In order to further understand the information content in the projected data, the first four PC scores are now used as features for unsupervised clustering in subsequent analyses. At this point, one could note that a combination between K-means and PCA enables to generate and interpret K clusters on a 4-dimensional PC subspace rather than a 10-dimensional feature subspace.

The results of K-means based on the PC scores are presented in Fig. 5, in which panels A and B represent different combinations of PC subspaces.

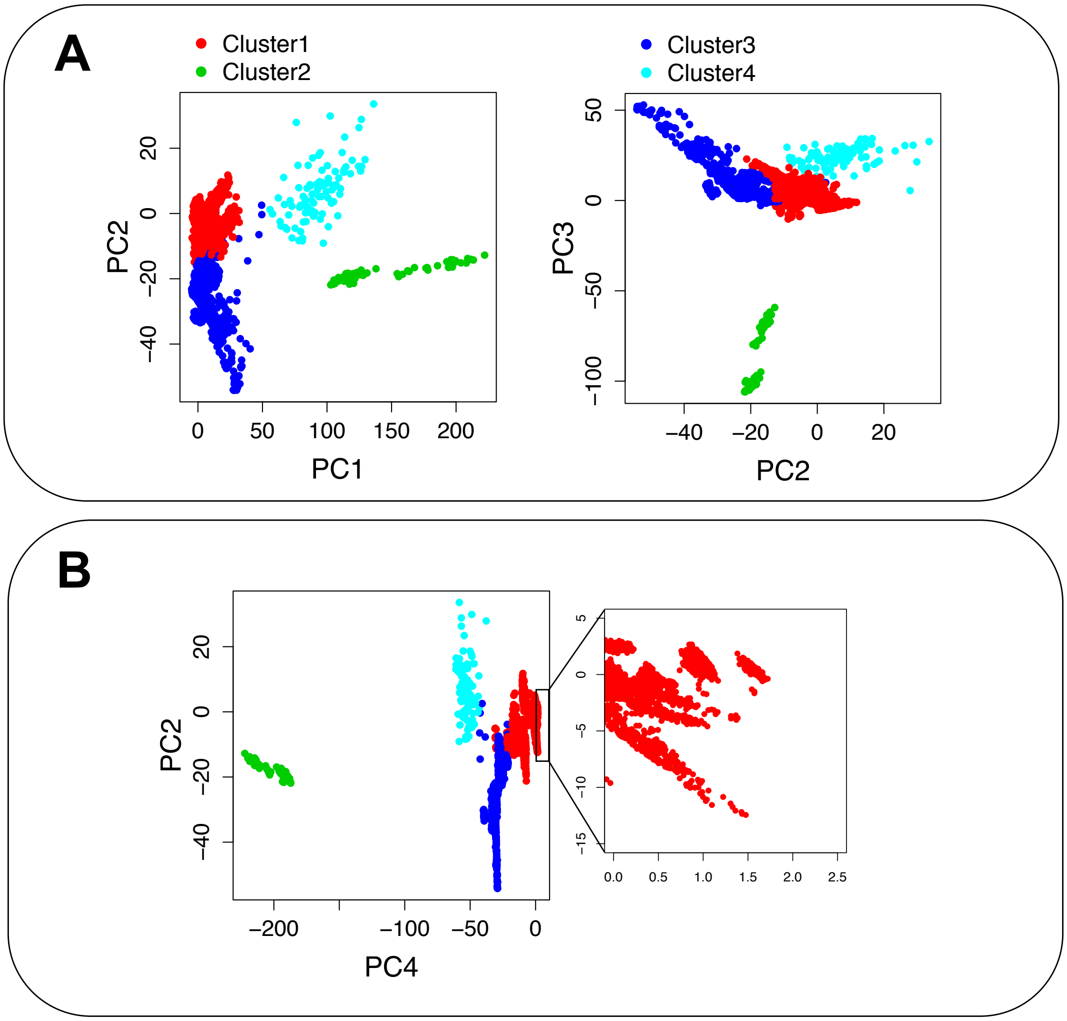

Fig. 5: Sample clusters determined by K-means. Panels A and B represent different combinations of PC subspaces.

A descriptive analysis of the variables in each cluster appears summarized in Table 3. The dynamics and location of the clustered samples on the different PC subspaces provide useful information concerning the behavior of the different types of components. In particular:

- All of the samples classified into Cluster 2 tend to assume highly positive values on the PC_1 (Panel A Fig. 5), particularly indicating that inventory levels for this component typology tend to be well-above average.
- The majority of the samples within Cluster 3 tend to strongly assume negative values over PC_2, demonstrating that the safety time, supply lead time and monetary value of stock on-hand for this class of components are above average. At this point, since safety time pushes delivery orders earlier, the larger the value of this parameter the greater the amount of stock on-hand and holding costs.

Thus, attending to the high average stock levels recorded for components within Cluster 3, company managers should analyze the possibility of decreasing the safety time parameter for some components within this cluster. This reduction is particularly relevant in the automotive industry in which carrying the lowest possible level of inventory without neglecting service level is a primary concern [52].

- We found that all samples in the Cluster 2 are also plotted in the negative direction of PC_3, thus particularly suggesting that the values of agreed MOQs with suppliers are well-above average for commodities. This opens a space so that the company can negotiate less MOQs with suppliers in order to decrease the high average stock levels related to this component typology (see Table 3).
- The Cluster 1 is the only one containing samples located in the positive direction of PC_4 (Panel B of Fig. 5). Concretely, 35% of the samples therein contained satisfy that condition. This suggests that, in general, average supplier on-time delivery (OTD) scores for some inventory components within this category tend to be higher than those recorded for components classified into the remaining clusters.

Table 3: Descriptive statistics of the features for different clusters.

Feature	Cluster 1 ($n = 8257$)		Cluster 2 ($n = 77$)		Cluster 3 ($n = 1376$)		Cluster 4 ($n = 96$)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
qty.rec.	2811.96	3644.45	3.52	1.67	4247.35	5050.92	62133.88	24589.42
saf.time	2.94	1.87	3.03	0.23	5.14	2.81	4.07	0.62
val.stock	23827.93	42600.12	13249.77	16386.74	69987.24	132501.93	49972.12	11700.65
cons.stock	1286.56	1384.23	6857.42	3152.33	669.99	870.99	9109.10	3211.48
supp.otd.	80.11	24.26	64.68	19.70	52.52	20.48	52.21	8.34
wh.occup.	11.57	11.13	0.00	0.00	6.69	7.27	3.02	1.89
stock	8766.68	10722.34	372110.78	212060.78	17120.40	19482.01	367195.32	85015.91
moq	678.51	1190.89	70784.42	1818.26	953.67	965.59	2541.46	562.97
supp.lt	3.96	3.99	4.49	3.73	26.40	4.77	26.82	1.74
nr.end	23.82	29.64	22.40	20.05	9.68	14.57	3.97	0.31

SD = Standard Deviation.

4.2 Gaining insights from clustered data

Aiming to get a more comprehensive knowledge of the clustering results, we analyzed the proportion of samples of each one of the 6 categories in the different clusters. These results are presented in Table 4, in which the component categories with a strong presence in each cluster are highlighted in boldface.

For the concerned company one of the core and most critical procedures is the shipment process to the end-customers. Therefore, company managers have suggested to analyze the dynamics of the different clusters according to two important variables, namely the average supplier OTD score and the total number of end-items that require a given component to be produced (Fig. 6).

Table 4: Distribution of categories within the four clusters derived by 4-means.

Category	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	n	%	n	%	n	%	n	%
High runner	**4818**	**58.4%**	0	0%	0	0.0%	0	0%
Stable	752	9.1%	0	0%	181	13.2%	1	1%
Special freights	382	4.6%	0	0%	**820**	**59.6%**	0	0%
Commodity	500	6.0%	**77**	**100%**	0	0.0%	0	0%
Common among plants	484	5.9%	0	0%	372	27.0%	**95**	**99%**
Critical	1321	16.0%	0	0%	3	0.2%	0	0%

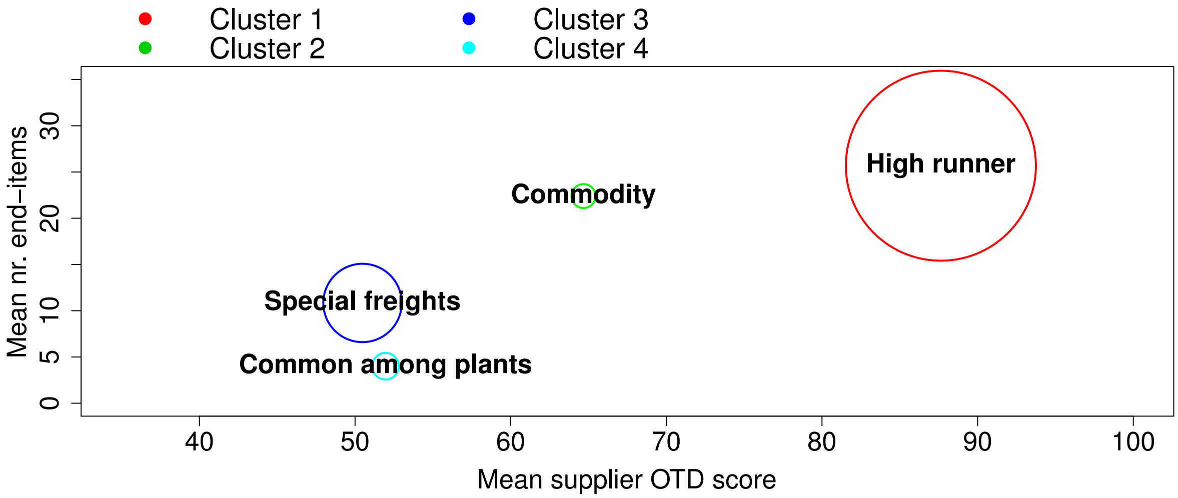

Fig. 6: Cluster dynamics according to selected logistic metrics.

In Fig. 6, each cluster traduces average values and is represented by a circle with radius proportional to the number of samples (n) of the concerned component category in that cluster. Moreover, all clusters are labelled according to the category that represents more than 50% of the total cluster size (see Table 4). We observe that high runner components (Cluster 1) are the ones with the highest average supplier OTD score. Furthermore, they are necessary to produce several end-items. By contrast, components prone to special freights show the smallest average supplier OTD score. This finding might seem contradictory at first inasmuch as one of the primary reasons of using a special freight is to avoid delays [53]. However, since special freights are last minute emergency shipments, just-in-time arrivals could be undermined if there is no timely detection for establishing the need for carrying out these shipments by the logistics planners. In such situations, premium freights are carried out but not sufficient to avoid time deviations from due dates or even production line stoppages, if the concerned components are necessary to produce several end-items as these ones are. Hence, as special freights are very costly, future requests should be carefully and timely planned.

In order to get a better insight into the environment in which the concerned company operates, we investigated the geographical distribution of the company suppliers according to the obtained clusters (Fig. 7). In this context, the cluster analyses enable us to ascertain that the majority of high runner component suppliers are located in Europe, in the neighbourhood of the concerned company. Conversely, components prone to special freights, which in turn present a lower average supplier OTD score, are mainly provided by Asian suppliers, normally associated with higher supply lead time values.

Following a subjective cluster evaluation, the aforementioned results and analyses derived therefrom were validated by the company managers, who confirmed the usefulness of the proposed approaches to enhance future decision making processes in the field of inventory management. For example, the strategies herein presented can bring relevant guidelines to set new parameter values into Enterprise Resource Planning (ERP) systems for the different components, that so far are established based on objective data analyses rather than technique. Furthermore, the visibility of the multiple components with multiple suppliers could also be enhanced with the adoption of these unsupervised learning techniques, enabling for instance to detect inventory target deviations. At the end, company managers would start adopting proactive behaviors rather than reactive ones. Another significant advantage resulting from the use of the proposed approach in practice is the possibility to enhance demand forecasting. Indeed, the classification of the samples into several homogeneous clusters allows to develop machine learning methods that are suitable for multiple (but similar) time series rather than train several models, one for each time series.

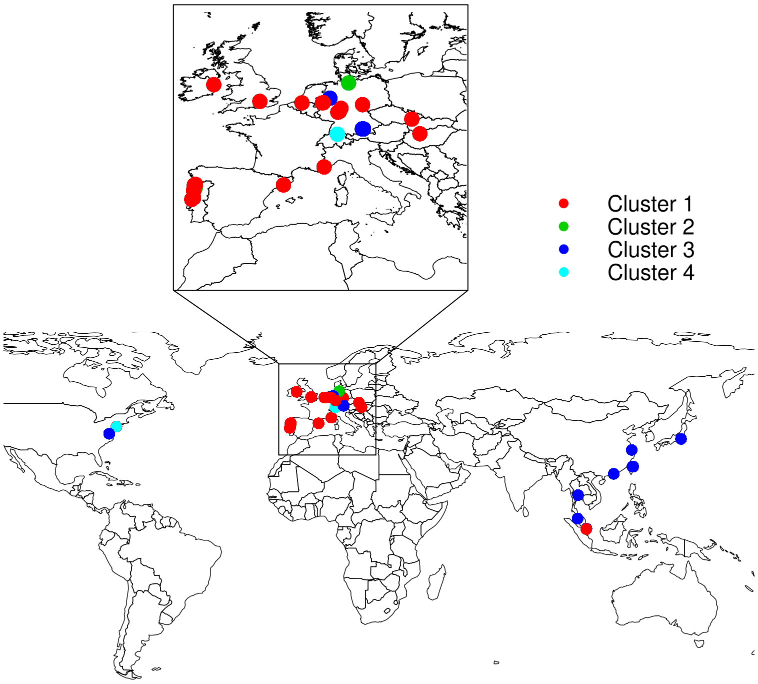

Fig. 7: Geographical distribution of the company suppliers according to the 4-means clustering.

4.3 Cluster validity

We compared the results obtained using K-means clustering with those using two flexible clustering algorithms: spherical K-means clustering [54] and spectral clustering [55].

The spherical K-means clustering is a variant of the classical K-means suitable for high dimensional datasets, which takes advantage of the cosine dissimilarity measure rather than the Euclidean metric. On the other hand, given a set of n p-dimensional data points $x_1, x_2, \ldots, x_n$, the classical spectral clustering transforms the raw data information into an *affinity graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node of $\mathcal{V}$ represents a particular data point while each edge of $\mathcal{E}$ traduces the similarity between two distinct data points. For each edge $(i, j) \in \mathcal{E}$, there is an associated weight w_{ij} that encodes the similarity (or affinity) between two data points x_i and x_j. We denote by $\mathcal{W} = (w_{ij})_{i,j=1}^{n}$ the *affinity matrix* of $\mathcal{G}$. Then, the ultimate goal is to partition $\mathcal{V}$ into K subsets $\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$. Nevertheless, since the classical spectral clustering generally has a computational complexity of $O(n^3)$, as a result from the computation of the eigenvectors of the $n \times n$ affinity matrix $\mathcal{W}$, its applicability to large-scale datasets becomes limited. For this reason, we adopt the Fast Approximate Spectral Clustering (FASP) algorithm [56], with gaussian mixture modeling (GMM) to reduce the high computation cost inherent to the classical spectral clustering algorithm.

To obtain the optimal number of clusters K for the spherical K-means clustering algorithm, we iterated it for K varying from 2 to 25 centers and compare the respective RS indexes. For the case of FASP algorithm, we follow a recent approach based on eigenvector distributional analysis proposed in [57]. As a result, we set $K = 4$, for both spherical K-means and FASP. Aiming to measure the quality of clustering results, namely in what concerns the compactness and separation of clusters, the *silhouette width method* [58] and a *Generalized Dunn's index* (GDI) [59] were employed using the Euclidean metric. The latter metric overcomes some drawbacks of the original Dunn's index (see [59] for details). The Generalized Dunn's index herein presented represents the ratio between the minimum average dissimilarity between two clusters and the maximum average dissimilarity within clusters. The higher the GDI, the better is the clustering. Regarding the silhouette width metric, it takes values in the compact interval $[-1, 1]$. For a given observation i, a value of $S(i)$ close to 1 translates into a good clustered observation (perfectly clustered for $S(i) = 1$). Conversely, a value $S(i)$ close to -1 indicates that i is probably a misclassified observation. In terms of internal cluster validation, we followed a 10-fold cross-validation approach to compute both silhouette and GDI metrics for the test set, by taking advantage of the **fpc R** package. Concretely, for each fold, each of the three clustering algorithms was applied to both train and test data. Then, the training centroids were used to classify the test observations into different clusters. The derived clusters were then validated according to the two distance based metrics previously described.

Table 5 presents the clustering evaluation results for the different validation methods used under 10-fold cross-validation. For this particular dataset, the results show that K-means generate reasonable structured clusters, outperforming the remaining clustering algorithms in terms of both considered cluster validation methods. In particular, when the Silhouette width is taken into consideration, the improvement rate of K-means is observed as 11.3% and 45.3% over spherical K-means and FASP, respectively. The superiority of K-means also holds when the GDI method is considered, leading to improvement rates of 42.6% and 38.3% over spherical K-means and FASP, respectively.

Table 5: Clustering evaluation results under 10-fold cross-validation for $K = 4$ (best mean values are highlighted in boldface).

	K-means		Spherical K-means		FASP	
Cluster validation method	Mean	SD	Mean	SD	Mean	SD
Silhouette width	**0.6839**	0.0107	0.6066	0.0950	0.3742	0.2889
Generalized Dunn's index	**0.8098**	0.1627	0.4652	0.3066	0.4998	0.4097

SD = Standard Deviation.

5 Conclusions

Understanding supply chain dynamics is a crucial task, especially with regard to inventory management. Motivated by the permanent pressure facing the automotive industry to meet customer orders whilst maintaining low inventory levels, we apply descriptive data mining techniques for profiling different inventory component categories. Concretely, we take advantage of real-world data collected from an automotive electronics SC to: (i) explore the application of PCA as a dimensional reduction technique in order to summarize the overall data structure, (ii) assess the relevance of combining partitional clustering with PCA to improve the extraction of important logistic information contained in the leading principle components, and (iii) provide some managerial guidelines to practitioners who intend to leverage inventory management for superior performance.

For the case study at stake, our findings suggest that further interpretation of the PCA results is hampered by the fact that several data samples from distinct component categories overlap at specific coordinates of the PC score plot. Thus, if the purpose is to identify relevant logistic patterns between the distinct component samples, partitional clustering is our preferred approach. Yet, when the PC scores are used as an input for clustering the task of profiling components according to the location of the respective clusters on the different PC subspaces is enhanced. Also, PCA revealed to be helpful in transforming our data into a lower dimensional representation rather than interpreting a higher-dimensional subspace. Thus, we argue in favor of adopting PCA in combination with K-means.

On the other hand, our results show that there is no relevant distinction among some component categories since they are classified into the same cluster. This therefore provides evidence in favor of the application of K-means to identify major clusters of similar components rather than, in practice, classify them in a manually fashion without multivariate information. The obtained clusters were subsequently validated via the average silhouette and generalized Dunn's indexes under 10-fold cross-validation. Overall, our results confirmed the benefits inherent to the application of unsupervised learning techniques for inventory components profiling in a real-world context. If applied, these methods have the potential to extract important insights from the data that may turn out to be very useful to enhance decision making processes related to the definition of suitable procurement strategies and inventory policies able to create value added in the supply chain while reducing supply chain costs. Yet, we acknowledge that the investigated methods should not be understood as a panacea to tackle any inventory management problem, but as a complementary tool with the ability to create value in dynamic supply chains. As future research, we intend to explore a wider set of explanatory variables, as well as to test different clustering algorithms under ensemble and consensus methods to derive better data partitions.

References

1. Areti Manataki, Yun-Heh Chen-Burger, and Michael Rovatsos. Scolog: A logic-based approach to analysing supply chain operation dynamics. *Expert Systems with Applications*, 41(1):23–38, 2014.
2. Karen Butner. The smarter supply chain of the future. *Strategy & Leadership*, 38(1):22–31, 2010.
3. Luciano Ferreira and Denis Borenstein. A fuzzy-bayesian model for supplier selection. *Expert Systems with Applications*, 39(9):7834–7844, 2012.
4. Douglas M Lambert and Martha C Cooper. Issues in supply chain management. *Industrial marketing management*, 29(1):65–83, 2000.
5. Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.
6. David Bendig, Malte Brettel, and Benedikt Downar. Inventory component volatility and its relation to returns. *International Journal of Production Economics*, 200:37–49, 2018.
7. Srinivas Talluri and Joseph Sarkis. A model for performance monitoring of suppliers. *International Journal of Production Research*, 40(16):4257–4269, 2002.
8. Fabian Dunke, Iris Heckmann, Stefan Nickel, and Francisco Saldanha-da Gama. Time traps in supply chains: Is optimal still good enough? *European Journal of Operational Research*, 264(3):813–829, 2018.
9. Bernhard Roßmann, Angelo Canzaniello, Heiko von der Gracht, and Evi Hartmann. The future and social impact of big data analytics in supply chain management: Results from a delphi study. *Technological Forecasting and Social Change*, 130:135–149, 2018.
10. Global Market Insights, Inc. Automotive electronics market - 8% growth forecast over 2017-2024. http://markets.businessinsider.com/news/stocks/Automotive-Electronics-Market-8-Growth-Forecast-over-2017-2024-1011544708. Accessed 14th January, 2018, 2017.
11. Alberto Petroni and Marcello Braglia. Vendor selection using principal component analysis. *Journal of supply chain management*, 36(1):63–69, 2000.
12. Rainer Lasch and Christian G Janker. Supplier selection and controlling using multivariate analysis. *International Journal of Physical Distribution & Logistics Management*, 35(6):409–425, 2005.
13. Charles V Trappey, Amy JC Trappey, Ai-Che Chang, and Ashley YL Huang. Clustering analysis prioritization of automobile logistics services. *Industrial Management & Data Systems*, 110(5):731–743, 2010.
14. You-Shyang Chen, Ching-Hsue Cheng, and Chien-Jung Lai. Extracting performance rules of suppliers in the manufacturing industry: an empirical study. *Journal of Intelligent Manufacturing*, 23(5):2037–2045, 2012.
15. UC Moharana and SP Sarmah. Joint replenishment of associated spare parts using clustering approach. *International Journal of Advanced Manufacturing Technology*, 94(5-8):2535–2549, 2018.
16. Chieh-Yuan Tsai, Chi-Yang Tsai, and Po-Wen Huang. An association clustering algorithm for can-order policies in the joint replenishment problem. *International Journal of Production Economics*, 117(1):30–41, 2009.
17. Faisal Aqlan. Dynamic clustering of inventory parts to enhance warehouse management. *European Journal of Industrial Engineering*, 11(4):469–485, 2017.
18. Hasan Kartal, Asil Oztekin, Angappa Gunasekaran, and Ferhan Cebi. An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering*, 101:599–613, 2016.
19. Francesco Lolli, Elia Balugani, Alessio Ishizaka, Rita Gamberini, Bianca Rimini, and A Regattieri. Machine learning for multi-criteria inventory classification applied to intermittent demand. *Production Planning & Control*, pages 1–14, 2018.
20. Cristopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
21. Jiawei Han, Jian Pei, and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2011.
22. Ian Jolliffe. *Principal Component Analysis*. New York: Springer, 2002.
23. Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
24. Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

25. Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
26. Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.
27. Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.
28. Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1(804):801, 1956.
29. Geoffrey H Ball and David J Hall. Isodata, a novel method of data analysis and pattern classification. Technical report, Stanford research inst Menlo Park CA, 1965.
30. James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
31. Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
32. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA, 2001.
33. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.
34. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
35. Subhash Sharma. *Applied multivariate techniques*. Wiley New York, 1996.
36. Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.
37. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
38. Forghani, Athena and Sadjadi, Seyed Jafar and Moghadam, Babak Farhang. A supplier selection model in pharmaceutical supply chain using PCA, Z-TOPSIS and MILP: A case study. *PLOS ONE*, 13(8):e0201604, 2018.
39. Park, Yong Shin and Egilmez, Gokhan and Kucukvar, Murat. A novel life cycle-based principal component analysis framework for eco-efficiency analysis: case of the United States manufacturing and transportation nexus. *Journal of Cleaner Production*, 92:327–342, 2015.
40. Park, Yang-Byung and Yoon, Sung-Joon and Yoo, Jun-Su. Development of a knowledge-based intelligent decision support system for operational risk management of global supply chains. *European Journal of Industrial Engineering*, 12(1):93–115, 2018.
41. Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.
42. Catherine Combes and Jean Azema. Clustering using principal component analysis applied to autonomy–disability of elderly people. *Decision Support Systems*, 55(2):578–586, 2013.
43. Andreas Adolfsson, Margareta Ackerman, and Naomi C Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, 2019.
44. Christophe Croux and Gentiane Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.
45. Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
46. Henry F Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151, 1960.
47. Jorge Cadima and Ian T Jolliffe. Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics*, 22(2):203–214, 1995.
48. Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC press, 2013.
49. Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
50. Glenn W Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342, 1980.
51. Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

52. Sherif A Masoud and Scott J Mason. Integrated cost optimization in a two-stage, automotive supply chain. *Computers & Operations Research*, 67:1–11, 2016.
53. Mualla Gonca Avci and Hasan Selim. A multi-objective simulation-based optimization approach for inventory replenishment problem with premium freights in convergent supply chains. *Omega*, 80:153–165, 2018.
54. Christian Buchta, Martin Kober, Ingo Feinerer, and Kurt Hornik. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012.
55. Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
56. Donghui Yan, Ling Huang, and Michael I Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916. ACM, 2009.
57. Christopher R John, David Watson, Michael Barnes, Costantino Pitzalis, and Myles J Lewis. Spectrum: Fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, pages 1–8, 2019.
58. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
59. James C Bezdek and Nikhil R Pal. Some new indices of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 28(3):301–315, 1998.