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

João N.C. Gonçalves[a,*], Paulo Cortez[b], M. Sameiro Carvalho[a]

[a]ALGORITMI Research Centre, Department of Production and Systems, University of Minho, 4710–057 Braga, Portugal
[b]ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4804–533 Guimarães, Portugal

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 SC. 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 improved SC 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 [19, 42], especially at a time where there is a pressing need for SC integration and coordination [7, 12, 26]. Bearing in mind that organizations are commonly structured in SCs [20], SC

[*] Correspondence: João N.C. Gonçalves (e-mail address: jncostagoncalves@gmail.com)

management (SCM) plays a paramount role in promoting their success, achieving their objectives and, above all, guaranteeing customer satisfaction [36, 38]. 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 [23]. In the literature, this process is closely bound up with the volatility of inventory components [8]. 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. This potentially leads to excess inventory, as well as to increased holding costs and higher risks of product obsolescence. It is, therefore, essential to strike the proper balance between stock-outs and overstock so that customer service levels are 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 logistic parameters associated with it, might provide important insights on how to manage it. On the other hand, since the inventory of components is directly influenced by interactions with suppliers, the buyer-supplier relationship can also be enhanced during this profiling process [58]. Yet, although 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 [18]. This, together with the complex business environments characterized by the rapid growth of generated data [52], 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 such techniques is particularly 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 [24]. Increasingly, data mining (DM) approaches [25] have been proposed to improve SC processes, for instance relating to the ranking, selection and evaluation of suppliers [50, 37, 60, 14]). Up until now, however, the application of DM techniques in the field of inventory management has not been fully explored [45]. For example, [61] introduced 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. [4]

applied K-means clustering to group inventory parts according to different features. The obtained clusters served as a guideline for warehouse space optimization. [35] considered 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 work of [40].

Focusing on methods that do not require a-priori knowledge of underlying patterns, also known as unsupervised methods [10], 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. In particular, the mathematical relationship between ten quantitative and dependent variables is firstly studied by taking advantage of classical theory 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 further 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 strategies in the field inventory management, already pointed out by [45]. 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 data collection procedure, as well as the selected unsupervised learning models. In Section 3 we describe the modeling framework. Afterwards, 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. 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. 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	Std.Dev.	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 a component 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	–	Component 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"})

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 components with high marginal propensity to incur in a premium freight (motivated by stock-out 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 naturally contain components belonging to more than one category.

2.2. Selected unsupervised learning models

Two unsupervised learning methods, namely 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 [31, 28] and references cited therein for details).

2.2.1. Principal Component Analysis

As a descriptive and multivariate statistical technique, PCA was firstly studied by [49] and [27]. PCA intends to compress the dimension of a given dataset, whilst minimizing statistical information loss [32]. Let $\mathbf{X}$ be an $n \times p$ data matrix with $\text{rank}(\mathbf{X}) = r$, containing p features on n observations. PCA sequentially finds unit vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ that maximize $\text{var}(\mathbf{Xv})$ with the additional constraint that $\mathbf{v_{i+1}}$ is orthogonal to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_i}$. These vectors are known as PC loadings, whereas the $\mathbf{Xv_i}$'s are the corresponding PCs [56]. The PCs are thus linear combinations of the original features and are uncorrelated with each other in a descending order of relevance in terms of total variance explained [1]. Each component can be then interpreted according to the inter-correlated variables that comprise it. 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 [1]. Typically, two approaches are often used to select the number of PCs to retain. The first one consists in selecting the PCs whose eigenvalues are larger than 1 [34]. 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 [31].

2.2.2. K-means clustering

Driven by the studies of [57, 6, 41, 39], K-means is an iterative descent clustering method [22], considered to be the most widely used algorithm for partitional clustering [64]. 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 [29], 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 this optimization problem be NP-hard, 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 [28]; (2) the distance metric considered – typically the Euclidean, notwithstanding other distance metrics can be used (e.g., Mahalanobis and Gower); and (3) the initial cluster assignment, also called cluster initialization. Regarding the third parameter, 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. Then, it is chosen the solution for which the optimization problem (1) is minimized [28, 29]. In this work, the optimal number K is selected via the R-squared (RS) [55] and the prediction strength [59] 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 $[0, 1]$. If the value of RS is 0, then 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 [59] for details).

3. Modeling framework

The numerical experiments presented throughout this section were conducted in the **R** programming language [51] with suitably selected packages.

Firstly, we adopted PCA in order to transform a set of correlated variables into a smaller set of linearly uncorrelated variables, which retain the most relevant information from the original dataset whilst minimizing information lost. With the application of PCA we intended 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 [21] or on the extraction of the most relevant sustainability indicators to conduct eco-efficiency performance analyses in industrial companies [48]. PCA can also contribute to the identification of operational risk sources [see, e.g., 47] 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 management decision-making process.

Secondly, the PC scores are used as input features for K-means clustering. Note that, following this approach, we intend 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 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 [17, 15, 2]. 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. Thus, to treat all input features on an equal basis, we preferred the use of the correlation matrix rather than the covariance matrix. 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 [32]. Since classical PCA is not robust to outliers and noise data, we further considered a Minimum Covariance Determinant (MCD)-based PCA [16]. The MCD method adopts a highly robust estimator of multivariate locator and scatter and has been explored to develop robust multivariate approaches [54]. Following this strategy, 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 [16]. Concerning the selection of the number of PC to be included, there exists a trade-off between increasing variance explained while reducing the number of PCs containing noise. Following common yet subjective practice [32], 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). Under the Kaiser's rule [33], the exclusion of the remaining PCs can be justified by the fact that the respective eigenvalues are not equal to or greater than 1. We additionally found through background analyses (not presented) that the different samples showed a strong overlap on the higher-order PCs, 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 PCs would essentially represent noise.

For the sake of interpretability, since PCs are linear combinations of all the dataset features, we identified, for each PC, which features can be discarded while preserving as much as possible statistical information. Typically, this identification is based on the magnitude of the feature loadings, neglecting those with low magnitude, which can be potentially misleading [13]. In a bid to reduce the subjective nature inherent to the interpretation of PCs, 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 [1].

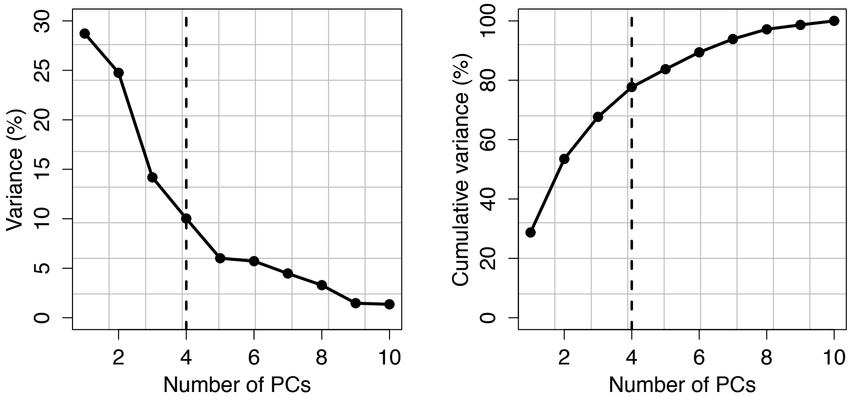

Figure 1: Variance (left) and cumulative variance (right) explained as a function of the number of PCs.

Figure 2 plots the correlation circles for the first four PC dimensions. 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 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 (PC1) 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 F_1, these features seem to have no strong correlation with PC2, which essentially contrasts the safety time (F_2) and the supplier lead time (F_9) with both the stock quantity received (F_1) by the organization and the number of end-items that make use of the component 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 allowed us to obtain truncated PCs ($\mathrm{PC}i^{tr}, i = 1, \ldots, 4$). Each $\mathrm{PC}i^{tr}$ is defined as follows:

$$\mathrm{PC1}^{tr} = 0.3983F_1 + 0.3288F_4 + 0.4491F_6 + 0.5126F_7 + 0.2684F_8 \tag{2}$$

$$\mathrm{PC2}^{tr} = 0.3243F_1 - 0.3771F_2 - 0.4338F_9 \tag{3}$$

$$\mathrm{PC3}^{tr} = 0.4953F_3 - 0.3836F_8 \tag{4}$$

$$\mathrm{PC4}^{tr} = 0.0577F_3 + 0.4485F_5 - 0.6148F_8 - 0.5702F_9 \tag{5}$$

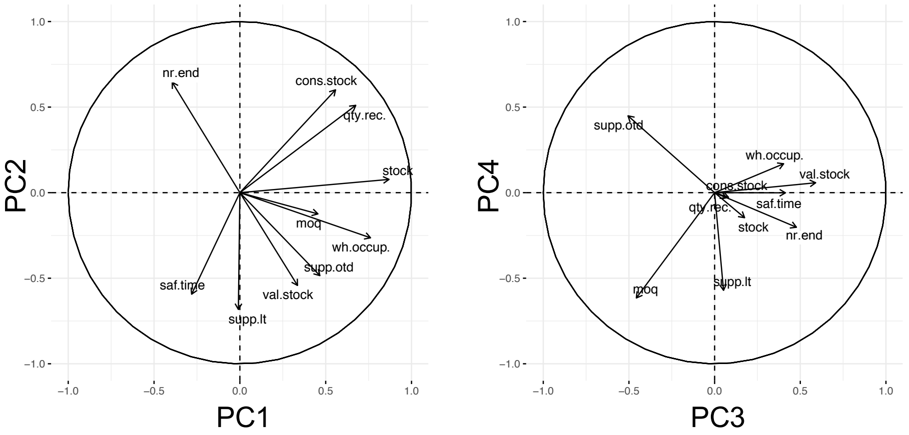

Figure 2: Correlation circles for the first and second (left), and third and fourth (right) PCs.

The selected subsets of features and the interpretation of each PCi^{tr} appear summarized in Table 2. Note that the interpretation of each PCi^{tr} depends both on the magnitude and signs of the variable loadings. For instance, the algebraic formulation of PC1^{tr} given by Eq. (2) shows that all the variables that comprise it are inventory-related and the corresponding loadings are positive. Thus, this suggests that PC1^{tr} can be interpreted as a weighted average of the inventory level, where samples with high PC1^{tr} scores exhibit high inventory levels, and vice versa. In contrast, PC2^{tr} comprises two variables with negative loadings and one variable with positive loading. Therefore, high values of PC2^{tr} reflect the contrast of the stock quantity received with the safety time and supply lead time. 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. In addition, whatever the truncated PC concerned, its correlation with the original PC is quite reasonable (≥ 0.8559), which corroborates the quality of approximation of the four extracted PCs using the truncated components.

Table 2: Summary of the truncated PCs.

PCi^{tr}	Subset of features	Corr(PCi^{tr}, PCi)	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
$i = 2$	$\{F_1, F_2, F_9\}$	0.8559	Contrast between F_1 and F_2, 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 [3]. However, apart from this strategy, there exist other initialization methods suitable for this purpose. In this work, 24 sets of cluster centers were obtained via the Ward's hierarchical agglomerative clustering method [63]. 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 [44, 3]. In this process, we considered the Euclidean metric and the Ward2 algorithm [46]. 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 was corroborated by the averaged prediction strength value (right of Fig. 3) attained for $K = 4$ ($\overline{ps}|_{K=4} = 0.8384$ with cutoff $= 0.8$ and 100 resampled datasets), which represents a proper threshold for obtaining well separated clusters [59].

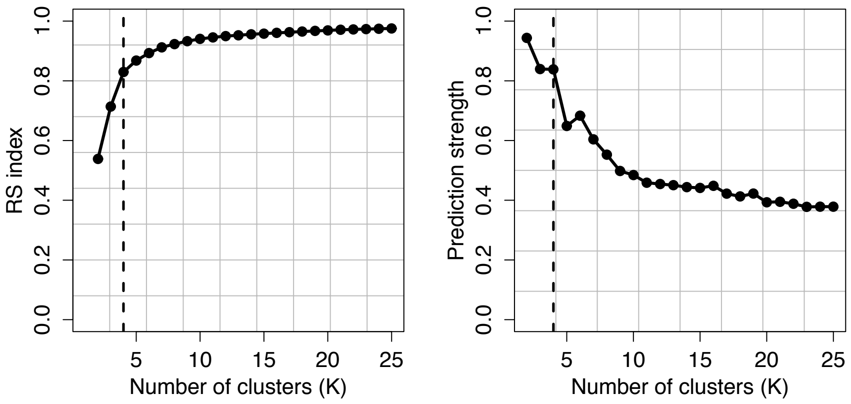

Figure 3: R-squared and prediction strength indexes as a function of K.

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 (pre-defined by the company managers).

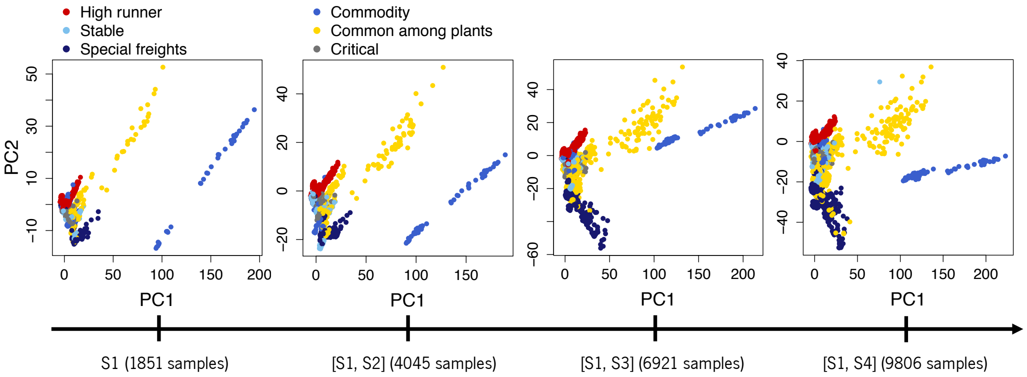

Figure 4: Evolution of the first two PC scores with increasing number of samples over four distinct time frames.

Since such categories are non-overlapping, we therefore coded each category 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 PC2, meaning that the averaged stock quantity received related to those components decreased substantially in that period. In addition, over the year 2016 ([S1, S2]), the samples of components prone to special freights were mainly located on the negative semi-axis of PC2, attaining minimum PC scores of close to -20. However, 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 PC2, which have translated into higher safety times and supply lead times for some materials that fall within that specific component category. Nevertheless, the overall samples showed a strong overlap on these two PC-dimensions, making it difficult to identify any further relevant information. 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 to study if this apparent drawback of PCA could be 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. Thus, in the following subsections, we use the PC scores as inputs for K-means to investigate if samples grouped into different component categories by the company managers are classified into the same cluster by K-means clustering.

4.1. Visualizing PC scores via K-means

The first four PC scores are now used as features for unsupervised clustering. The results of K-means based on the PC scores are presented in Fig. 5. Panels A and B represent different combinations of PC subspaces. A comparison between the sample distribution projected in the score plot of Fig. 4 with that observed in Fig. 5 reveals that the six component categories are now grouped into 4 different clusters. Thus, we may argue that the manually classification of some samples might have been conducted incorrectly by the company managers, in the sense that some similar samples (grouped into the same cluster) seem to have been previously classified into distinct component categories. 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 PC1 (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 PC2, demonstrating that both the safety time and supply lead time 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 averaged stock levels recorded for components within Cluster 3 (in terms of quantity and monetary value), company managers should analyze the possibility of improving demand forecasts for some components within this cluster in order to decrease the respective safety time parameters. 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 [43].

- We found that all samples in the Cluster 2 are plotted in the negative direction of PC3, 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 averaged 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 PC4 (Panel B of Fig. 5). Concretely, 35% of the samples therein contained satisfy that condition. This suggests that, in general, averaged 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.

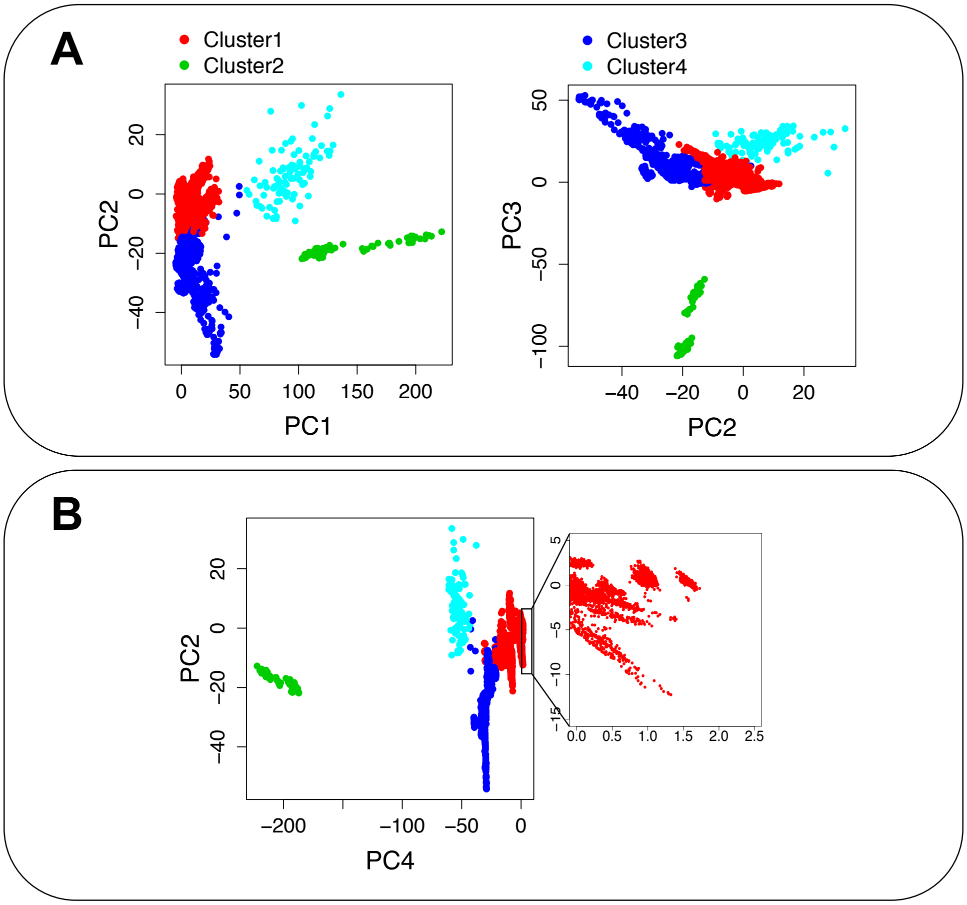

Figure 5: Sample clusters determined by K-means. Panels A and B represent different combinations of PC subspaces.

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	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
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

4.2. Gaining insights from clustered data

To further get an insight into the results, we also analyzed the proportion of samples of each one of the 6 categories in the different clusters (Table 4).

Table 4: Distribution of categories within the four clusters derived by 4-means. Component categories with a strong presence in each cluster are highlighted in boldface.

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%

For the concerned company one of the core and most critical procedures is the shipment process to the end-customers. Therefore, we analyzed the dynamics of the different clusters according to two important variables, namely the averaged supplier OTD score and the total number of end-items that require a given component to be produced (Fig. 6). In Fig. 6, each cluster traduces averaged 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). One can observe that high runner components (Cluster 1) are the ones with the highest averaged supplier OTD score. Furthermore, they are necessary to produce several end-items. Conversely, components prone to special freights show the smallest averaged 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 [5]. 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, special 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.

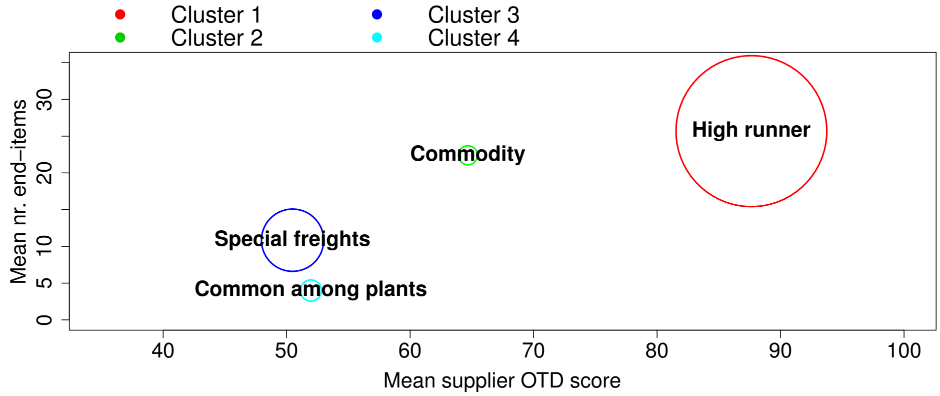

Figure 6: Cluster dynamics according to selected logistic metrics.

Concerning the geographical distribution of the company suppliers according to the obtained clusters (Fig. 7), we found that the majority of suppliers for high runner components are located in Europe, in the neighbourhood of the concerned company. On the other hand, components prone to special freights, which present a lower averaged supplier OTD score, are mainly provided by Asian suppliers, normally associated with higher supply lead times.

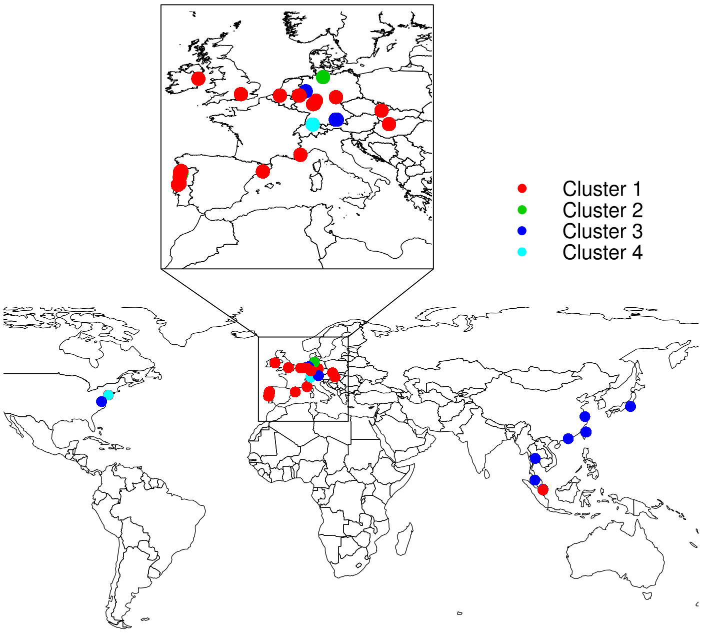

Figure 7: Geographical distribution of the company suppliers according to the 4-means clustering.

4.3. Cluster validity

We compared the results obtained using K-means clustering with those using two flexible clustering algorithms: spherical K-means clustering [11] and spectral clustering [62]. 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 $\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$. However, 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 adopted the Fast Approximate Spectral Clustering (FASP) algorithm [65], with gaussian mixture modeling (GMM) in order 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 compared the respective RS indexes. For the case of FASP algorithm, we follow a recent approach based on eigenvector distributional analysis proposed in [30]. As a result, we set $K = 4$, for both spherical K-means and FASP. To measure the quality of clustering results, namely in what concerns the compactness and separation of clusters, the silhouette width method [53] and a Generalized Dunn's index (GDI) [9] were employed using the Euclidean metric. 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. 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.

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	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
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

For this particular dataset, the results show that K-means generates reasonable structured clusters, outperforming the remaining clustering algorithms in terms of the considered cluster validation methods. In particular, when the Silhouette width is taken into consideration, the improvement rate of K-means is observed as 12.7% and 82.8%

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 74.1% and 62% over spherical K-means and FASP, respectively.

4.4. Practical and managerial implications

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 the detection of inventory target deviations. At the end, company managers would start adopting proactive behaviors rather than reactive ones. On the other hand, this offers the opportunity to develop flexible demand forecasting models to improve the management of critical components with deviant inventory performances. These improved forecasts provide aligned decision making regarding short to middle term inventory management operations in a data-driven fashion. Finally, the classification of the samples into several homogeneous clusters also enables the development of forecasting strategies suitable for multiple (but similar) time series rather than train several models, one for each time series.

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 applied descriptive data mining techniques for profiling different inventory component categories. Concretely, we took 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 improved SC 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 their location 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. Therefore, we argue in favor of adopting PCA in combination with K-means. Our results also provided 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 averaged silhouette and generalized Dunn's indexes under 10-fold cross-validation. Overall, this work evidenced the benefits inherent to the application of descriptive DM techniques for profiling inventory components in a real-world context. If applied, these approaches have the potential to extract important insights from the data that may turn out to be very useful to enhance decision making processes related, for instance, to the definition of suitable procurement strategies and inventory control policies. 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 improve supply chain management. 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.

Acknowledgments

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020. The authors want to extend grateful thanks to the both Editors and reviewers whose comments have greatly improved the quality of the paper.

[1] Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

[2] Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26.

[3] Aggarwal, C. C. and Reddy, C. K. (2013). *Data clustering: algorithms and applications.* CRC press.

[4] Aqlan, F. (2017). Dynamic clustering of inventory parts to enhance warehouse management. *European Journal of Industrial Engineering*, 11(4):469–485.

[5] Avci, M. G. and Selim, H. (2018). A multi-objective simulation-based optimization approach for inventory replenishment problem with premium freights in convergent supply chains. *Omega*, 80:153–165.

[6] Ball, G. H. and Hall, D. J. (1965). ISODATA, a novel method of data analysis and pattern classification. Technical report, Stanford research inst Menlo Park CA.

[7] Barve, A., Kanda, A., and Shankar, R. (2009). The role of human factors in agile supply chains. *European Journal of Industrial Engineering*, 3(1):2–20.

[8] Bendig, D., Brettel, M., and Downar, B. (2018). Inventory component volatility and its relation to returns. *International Journal of Production Economics*, 200:37–49.

[9] Bezdek, J. C. and Pal, N. R. (1998). Some new indices of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 28(3):301–315.

[10] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

[11] Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22.

[12] Butner, K. (2010). The smarter supply chain of the future. *Strategy & Leadership*, 38(1):22–31.

[13] Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics*, 22(2):203–214.

[14] Chen, Y.-S., Cheng, C.-H., and Lai, C.-J. (2012). Extracting performance rules of suppliers in the manufacturing industry: an empirical study. *Journal of Intelligent Manufacturing*, 23(5):2037–2045.

[15] Combes, C. and Azema, J. (2013). Clustering using principal component analysis applied to autonomy–disability of elderly people. *Decision Support Systems*, 55(2):578–586.

[16] Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618.

[17] Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM.

[18] Dunke, F., Heckmann, I., Nickel, S., and Saldanha-da Gama, F. (2018). Time traps in supply chains: Is optimal still good enough? *European Journal of Operational Research*, 264(3):813–829.

[19] Faisal, M. N., Banwet, D., and Shankar, R. (2007). Quantification of risk mitigation environment of supply chains using graph theory and matrix methods. *European Journal of Industrial Engineering*, 1(1):22–39.

[20] Ferreira, L. and Borenstein, D. (2012). A fuzzy-bayesian model for supplier selection. *Expert Systems with Applications*, 39(9):7834–7844.

[21] Forghani, A., Sadjadi, S. J., and Moghadam, B. F. (2018). A supplier selection model in pharmaceutical supply chain using PCA, Z-TOPSIS and MILP: A case study. *PLOS ONE*, 13(8):e0201604.

[22] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA.

[23] Giannoccaro, I. and Pontrandolfo, P. (2002). Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161.

[24] Global Market Insights, Inc. (2017). 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.

[25] Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

[26] He, R., Xiong, Z., and Xiong, Y. (2015). Supply chain collaboration with complementary quality design and greener production. *European Journal of Industrial Engineering*, 9(4):470–511.

[27] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

[28] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

[29] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[30] John, C. R., Watson, D., Barnes, M., Pitzalis, C., and Lewis, M. J. (2019). Spectrum: Fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, pages 1–8.

[31] Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer.

[32] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

[33] Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151.

[34] Kaiser, H. F. (1961). A note on guttman's lower bound for the number of common factors 1. *British Journal of Statistical Psychology*, 14(1):1–2.

[35] Kartal, H., Oztekin, A., Gunasekaran, A., and Cebi, F. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering*, 101:599–613.

[36] Lambert, D. M. and Cooper, M. C. (2000). Issues in supply chain management. *Industrial marketing management*, 29(1):65–83.

[37] Lasch, R. and Janker, C. G. (2005). Supplier selection and controlling using multivariate analysis. *International Journal of Physical Distribution & Logistics Management*, 35(6):409–425.

[38] Lehoux, N., D'Amours, S., and Langevin, A. (2010). A win-win collaboration approach for a two-echelon supply chain: a case study in the pulp and paper industry. *European Journal of Industrial Engineering*, 4(4):493–514.

[39] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[40] Lolli, F., Balugani, E., Ishizaka, A., Gamberini, R., Rimini, B., and Regattieri, A. (2018). Machine learning for multi-criteria inventory classification applied to intermittent demand. *Production Planning & Control*, pages 1–14.

[41] MacQueen, J. et al. (1967). 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.

[42] Manataki, A., Chen-Burger, Y.-H., and Rovatsos, M. (2014). Scolog: A logic-based approach to analysing supply chain operation dynamics. *Expert Systems with Applications*, 41(1):23–38.

[43] Masoud, S. A. and Mason, S. J. (2016). Integrated cost optimization in a two-stage, automotive supply chain. *Computers & Operations Research*, 67:1–11.

[44] Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342.

[45] Moharana, U. and Sarmah, S. (2018). Joint replenishment of associated spare parts using clustering approach. *International Journal of Advanced Manufacturing Technology*, 94(5-8):2535–2549.

[46] Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295.

[47] Park, Y.-B., Yoon, S.-J., and Yoo, J.-S. (2018). 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.

[48] Park, Y. S., Egilmez, G., and Kucukvar, M. (2015). 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.

[49] Pearson, K. (1901). 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.

[50] Petroni, A. and Braglia, M. (2000). Vendor selection using principal component analysis. *Journal of Supply Chain Management*, 36(1):63–69.

[51] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[52] Roßmann, B., Canzaniello, A., von der Gracht, H., and Hartmann, E. (2018). 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.

[53] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

[54] Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

[55] Sharma, S. (1996). *Applied multivariate techniques*. Wiley New York.

[56] Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034.

[57] Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1(804):801.

[58] Talluri, S. and Sarkis, J. (2002). A model for performance monitoring of suppliers. *International Journal of Production Research*, 40(16):4257–4269.

[59] Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528.

[60] Trappey, C. V., Trappey, A. J., Chang, A.-C., and Huang, A. Y. (2010). Clustering analysis prioritization of automobile logistics services. *Industrial Management & Data Systems*, 110(5):731–743.

[61] Tsai, C.-Y., Tsai, C.-Y., and Huang, P.-W. (2009). An association clustering algorithm for can-order policies in the joint replenishment problem. *International Journal of Production Economics*, 117(1):30–41.

[62] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

[63] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

[64] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

[65] Yan, D., Huang, L., and Jordan, M. I. (2009). Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916. ACM.