Predicting Material Requirements in the Automotive Industry Using Data Mining

Tobias Widmer(✉), Achim Klein, Philipp Wachter, and Sebastian Meyl

University of Hohenheim, Stuttgart, Germany
{tobias.widmer, achim.klein, philipp.wachter, sebastian.mey}@uni-hohenheim.de

Abstract. Advanced capabilities in artificial intelligence pave the way for improving the prediction of material requirements in automotive industry applications. Due to uncertainty of demand, it is essential to understand how historical data on customer orders can effectively be used to predict the quantities of parts with long lead times. For determining the accuracy of these predications, we propose a novel data mining technique. Our experimental evaluation uses a unique, real-world data set. Throughout the experiments, the proposed technique achieves high accuracy of up to 98%. Our research contributes to the emerging field of data-driven decision support in the automotive industry.

Keywords: Predictive manufacturing · Material requirements planning · Data mining · Artificial intelligence · Automotive industry

1 Introduction

Predictive manufacturing has become a major challenge to the industrial production sector [1]. Manufacturers integrate business information systems into their production environment to create competitive advantages and to enhance efficiency and productivity [2]. These information systems increasingly use artificial intelligence for planning and controlling manufacturing operations [3]. In particular, the automotive industry is progressively adopting methods from artificial intelligence research in a wide range of industrial applications. For instance, Audi has developed intelligent Big Data capabilities to optimize their production and sales processes [4].

Cars are subject of increasing individualization, which exemplifies in the high number of possible variants. This complexity puts a burden on supply chain management in the automotive industry and calls for intelligent business information systems, which integrate supply and demand [5, 6]. For instance, BMW offers more than 10^{32} car variants, out of which several thousands are in fact ordered by customers [7]. Due to the high variance in products given globally distributed production plants, car manufacturers use planned orders based on forecasts to optimize their material requirements planning. As manufacturers transit from build-to-stock to build-to-order strategies, planning processes are reorganized by implementing advanced planning systems such as predictive manufacturing. In particular, quantities of car parts with long lead times must

W. Abramowicz and R. Corchuelo (Eds.): BIS 2019, LNBIP 354, pp. 147–161, 2019.
https://doi.org/10.1007/978-3-030-20482-2_13

be predicted accurately to prevent shortages and excess stock, respectively. Predictive manufacturing systems provide tools and methods to process historical data about customized orders into information that can explain planning uncertainties and support managers in making more informed decisions. These decisions typically concern strategies for planning material requirements along the entire supply chain.

To facilitate an efficient production in the presence of long lead time suppliers, manufacturers depend on accurate estimates about the material requirements for production. The increasing number of available options and option combinations for vehicle equipment entails highly complex and interdependent parts requirements lists (PRL) that are necessary to build a vehicle. Due to uncertainties emerging from suppliers with long lead times, however, manufacturers do not know in advance the exact quantity of the parts and components needed at each production plant [8]. While manufacturers use historical customer orders to estimate future sales and analytical high-level models for production planning, they have not yet exploited the full potential of their data to predict fine-grained material requirements. Therefore, we contribute a technique that exploits a unique dataset of fully specified vehicle orders with all product options and required material parts for predicting the material requirements of parts with long lead times.

Given the incomplete vehicle specifications of estimated future customer orders, it is essential to understand how historical orders can be used more effectively to improve the prediction for parts with potentially long lead times. We aim to enhance this understanding by proposing a data mining technique for predicting the quantities of parts with long lead times. These parts must be ordered at a time where the associated customer orders are not yet available. Therefore, we base our prediction on historical customer orders. We represent these orders as vectors in which each element corresponds to the frequency of a product option ordered by a set of customers. First, we exploit the concept of cosine similarity for quantifying the similarity between orders of different customer sets [9]. Then, we select the most similar set of customer orders and use the associated set of known required parts as predicted parts for the estimated set of future orders. Finally, we quantify the prediction quality of our approach using accuracy defined as the ratio between the predicted quantity and the actual quantity of parts.

To validate our proposed technique, we carried out a set of experiments using a unique data set, covering real-world purchase orders placed at an international automobile manufacturer during a fixed period of time. These orders contain information about the specific combinations of product options ordered by customers and associated required material parts during a fixed production cycle.

We calculated the accuracy of the prediction for customer order groups of varying sizes and uncertainty levels. We find that larger customer order groups yield higher accuracy across different uncertainty levels. Specifically, we find that our proposed technique achieves an accuracy between 88% and 98% throughout all experiments. This finding is consistent with the growing trend toward modularization in the automotive production industry [10]. As digital technology platforms emerge in smart production environments, car manufacturers can begin exploiting digitalized infrastructures to design and control innovative components of higher levels of standardization [11]. Our findings help explain to what degree the increasing modularization and standardization of components impact the prediction of the requirements for parts with long lead times.

The remainder of this paper is organized as follows. The next section discusses prior research. In Sect. 3, we present the proposed data mining technique for predicting material requirements. In Sect. 4, we report on the experimental evaluation and discuss our findings. Finally, we provide our conclusion in Sect. 5.

2 Prior Research

We discuss prior research on (1) predicting material requirements in industrial production under demand uncertainty, and (2) data mining approaches for estimating the similarity between different sets of product orders.

2.1 Predicting Material Requirements

Manufacturing industries use material requirements planning (MRP) systems for inventory management and for planning and forecasting the quantities of parts and components required for production. Since their first implementation in the 1960s, MRP systems evolved into MRP II and later into enterprise resource planning (ERP) systems. MRP systems use master and transaction data as input. While master data include information about structure and variants of each component, transaction data are created when a customer places an order [12]. In the presence of demand uncertainty, planning systems typically follow either a supply-oriented or a demand-oriented approach.

First, in **supply-oriented approaches**, manufacturers estimate the required quantities by optimizing a given objective function subject to production capacity, storage, and market constraints [12]. Gupta and Maranas [13] develop a stochastic model for minimizing the total cost of a multi-product and multi-site supply chain under uncertain demand. They solve the objective function using optimization methods in two stages. First, all manufacturing decisions are made before the demand is known. Second, inventory levels, supply policies, safety stock deficits and customer shortages are determined after the demand is already known while taking the quantities produced in the first stage into account. The main difference to our approach is that Gupta and Maranas analytically solve a model on the level of different products while we use a fully data-based approach for predicting material parts requirements of products. Whereas Gupta and Maranas focus on total production and logistics cost, we implicitly optimize cost by predicting material requirements for reducing potential over- or underproduction.

Second, **demand-oriented approaches** for production planning focus on forecasting future demand and adjusting the production accordingly. Common techniques used for this purpose include moving averages and exponential smoothing based on historical customer orders [12]. Zorgdrager et al. [14] compare various regression and statistical models to forecast the material demand for aircraft non-routine maintenance. They find that the exponential moving average model offers the best tradeoff between forecast errors and robustness over time.

Lee et al. [15] model uncertain demand using fuzzy logic theory. They integrate triangular fuzzy numbers in a part-period balancing lot-sizing algorithm to determine the optimal lot size under uncertain demand. Chih-Ting Du and Wolfe [16] propose a

decision support system to determine the optimal ordering quantity for materials and the safety stock. The system utilizes fuzzy logic controllers and neural networks and takes variables such as the ordering costs, inventory carry costs and uncertainty into account. The role of the neural networks is to increase the fault tolerance of the system and increase rule evaluation performance by learning and replacing imprecise and complicated fuzzy if-then rules. In contrast to Lee et al., we model uncertainty of demand in terms of ordered vehicle configurations in a simpler way by randomly removing product options from individual orders in our dataset.

Steuer et al. [17] predict the total demand for new automotive spare parts in three steps. First, they cluster automotive parts with similar product life cycle curves using k-medoids clustering and chi-square distances. The optimal number of clusters is determined by calculating the Dunn index and the silhouette width of the clusters for various number of clusters. Second, their approach identifies common features that products in the same cluster share. Third, they use a classification model to match new parts to clusters by estimating the feature similarity between them. They find that among all evaluated algorithms, support vector machines achieve the highest accuracy of 68.4%. We extend the cluster method of Steuer et al. by integrating the different option combinations of the vehicles ordered by customers. Whereas Steuer et al. focuses on spare parts only, we consider the material requirements for the complete production process.

In summary, existing approaches do not internalize real-world customer orders that include all possible option configurations. We contribute a novel data mining technique for predicting fine-grained material parts requirements given uncertain demand about vehicle option configurations. To this end, our approach can be used to complement existing approaches for fine-tuning the prediction of material parts.

2.2 Data Mining Approaches

Data-based prediction of required material parts in a production supply chain relates to analyzing historical parts requirements in transaction data for a certain product demand pattern. As such, predicting parts requirements involves comparing imprecisely specified current demand with closest known demand pattern for which the required parts are known. Thus, we face two problems: (1) forming clusters in historical transaction data (containing product options and required parts) to model potential product demand with respect to certain product options, and (2) measuring closeness of imprecise current demand vs. historical demand patterns. Because we tackle both problems on the basis of the vector space model, we first motivate and review the model and specific approaches for the two outlined problems.

The vector space model is a well-established model in the fields of data mining and text mining [18]. This model has been widely used for pattern matching and in particular for text retrieval and text classification [19]. In the scope of this work, we are interested in comparing and matching patterns of imprecise current demand with fully specified historical demand, for which required parts are known. Thus, the quantities of order details (e.g., ten times product type A, five times product option B) are used as elements of a vector. While the vector-based approach received little attention for predicting parts requirements in prior research, we contribute to existing literature by

transferring previous findings from using the vector space model in text mining to demand prediction for car production.

A major challenge in interpreting demand patterns as vectors in vector space is the assumption of linear independence of the dimensions of a vector space. It seems obvious that the dimensions offered by product types and product options are not fully independent. However, the quantification of text as vectors by interpreting the words of the vocabulary as dimensions and counting word occurrences in a vector's elements also clearly violates the independence assumption. The reason is that words in a sentence or a full text depend on each other. The same is true for product options, when cars are configured by customers (e.g., demand for certain luxury product options might be correlated). Despite the obvious violation of independence, text mining research has shown that text classification approaches still achieve high accuracy [20]. These findings also apply to our research as reported in Sect. 4.2 and discussed in Sect. 4.3.

Another challenge includes the imprecise formulation of demand patterns that must be represented as vectors. Imprecise or uncertain knowledge means that some quantities might not be known exactly. Furthermore, the quantities for some options might be completely unknown. Thus, the corresponding vector elements are zero, which leads to sparsity in the vector. To this respect, text classification research provides evidence for high performance despite sparse vectors, referring to the application of Support Vector Machines [21]. Our data mining technique is different because it is based on cosine similarities at its core. However, our results indicate high predictive performance.

Several clustering approaches are available for addressing the problem of forming clusters of historical transactions to create synthetic demand patterns. A possible approach consists in selecting transaction data randomly to form a cluster. Another option is choosing transactions based on similarity. In this case clustering algorithms from the field of unsupervised machine learning can be used. A prominent example for such an algorithm is k-means clustering [22]. K-means clustering partitions transaction data in the vector space by iteratively forming k clusters around centroid vectors with a minimum sum of squared distances of all other vectors with respect to the centroid cluster. Apart from this algorithm, research in data mining examines approaches based on hierarchy, fuzzy theory, distribution, density, graph theory, grids, fractal theory, and other models [23].

A number of approaches for addressing the problem of closeness of vectors have been proposed. Examples include the Minkowski distance, Euclidian distance, cosine similarity, Pearson correlation distance, and the Mahalanobis distance [23]. In the context of text mining, the most common approach is the cosine similarity. Cosine similarity measures the angle between two vectors in the same vector space. Uncertain demand, represented as vector with elements counting product options, is compared to fully specified demand vectors by the angles between the vectors. It is then assumed that the demand vector with the smallest angle is the most similar to the uncertain demand vector.

We address both problems in vector space, i.e., forming clusters and measuring closeness. Thus, we evaluate the suitability and performance (i.e., prediction quality) of our approach and contribute to the transfer of knowledge in text-related research in information systems (e.g., [24]) for predicting material requirements.

3 Data Mining Technique

We describe our data mining technique for predicting material parts by providing formal notations, illustrating its use in an example, and defining an accuracy measure. The proposed technique predicts quantities for parts with long times based on historical customer orders.

3.1 Measuring Similarity of Customer Order Groups

A single customer order describes a fully customized vehicle as ordered by a customer (e.g., through a web-based car configurator). A typical customer order includes a car configuration such as car model, engine type, navigation system, electric exterior mirrors, sunroof, and so on. Each customer order is accompanied by a parts requirements list (PRL). This list is used by the production plant to assemble the vehicle.

Now suppose the manufacturer wants to predict the quantity of parts required to produce a set of cars X given a basic configuration of product options. To achieve this, all historical customer orders are divided into groups randomly. All groups are sized equally by a pre-determined size. Let $G = \{g_1, \ldots, g_n\}$ be the set of all customer order groups. Further, let $O = \{o_1, \ldots, o_m\}$ be the set of distinct options present in G and X. Each group and also X are then represented by an m-dimensional vector $\overrightarrow{o_g}$. Further, let $f(g, o)$ denote the frequency of option $o \in O$ in group $g \in G$. Then, the vector representation is given by

$$\overrightarrow{o_g} = (f(g, o_1), \ldots, f(g, o_m)) \tag{1}$$

The vector representation assumes linear independence of the dimensions, which may not hold, but still the approach has achieved good results in other fields [20]. Once the vectors for all groups are formed, we measure the similarity between the set X of cars to produce and a historic group of cars $g \in G$ by calculating the cosine of the associated angles. We use cosine similarity because of its simplicity and effectiveness to get an initial validation of our data mining technique. The cosine similarity between vectors $\overrightarrow{o_g}$ and $\overrightarrow{o_X}$ can be derived using the Euclidean dot product formula,

$$S\left(\overrightarrow{o_g}, \overrightarrow{o_X}\right) := \cos(\theta) = \frac{\overrightarrow{o_g} \cdot \overrightarrow{o_X}}{\overrightarrow{o_g} \cdot \overrightarrow{o_X}} \tag{2}$$

Because each dimension within the vectors $\overrightarrow{o_g}$ and $\overrightarrow{o_X}$ equals the frequency of a distinct option in the corresponding groups and these frequencies cannot be negative, the cosine similarity is bounded in the interval $[0, 1]$. Thus, the closer S gets to 1, the more similar are X and g. If $S = 1$, X and g are said to be identical. In other words, we use the required parts to produce cars in group $g \in G$ with highest associated similarity value regarding X as prediction for required parts in X.

3.2 Illustrative Example

We provide a simple example to illustrate our data mining technique for estimating the parts requirements based on historical customer orders. Suppose a manufacturer seeks to forecast the number of parts required for producing a set X of 10 cars out of which 8 will have option o_1 (e.g., navigation system) and 4 will have option o_2 (e.g., sunroof). All other options are unknown to the manufacturer at this point in time. Implicitly, the frequency of unknown options is assumed to be zero. Hence, the vector representation for set X is given by

$$\overrightarrow{o_X} = (f(X, o_1), f(X, o_2)) = (8, 4) \tag{3}$$

To predict the number of parts in the presence of uncertainty about the final configuration, we divide the complete set of customer orders into 10 random groups of size 10 respectively; that is, $G = \{g_1, \ldots, g_{10}\}$. Each of these groups contain a set of distinct options. For instance, suppose customer order group g_1 consists of 10 cars out of which 9 are configured with option o_1 (i.e., navigation system), 7 are configured with option o_2 (i.e., sunroof), and 3 are configured with option o_3 (e.g., electric exterior mirrors). The vector representation of this group is then given by

$$\overrightarrow{o_{g_1}} = (f(g_1, o_1), f(g_1, o_2), f(g_1, o_3)) = (9, 7, 3) \tag{4}$$

Likewise, all other groups $\{g_2, \ldots, g_{10}\}$ are represented by vectors containing the frequency of product options over all orders of a group. Using cosine similarity, our data mining technique now discovers the group that is most similar to set X:

$$S\left(\overrightarrow{o_{g_1}}, \overrightarrow{o_X}\right) = \frac{(9, 7, 3) \cdot (8, 4, 0)}{(9, 7, 3) \cdot (8, 4, 0)} \approx 0.9483 \tag{5}$$

The cosine similarity between set X and the remaining groups $\{g_2, \ldots, g_{10}\}$ is calculated analogously. Suppose group g_1 is closest to X according to cosine similarity; that is, among all cosine similarities, 0.9483 is closest to 1. Thus, we use PRL of group g_1 as prediction for the PRL of set X.

3.3 Measuring the Accuracy of Predictions of Parts Requirements

Each individual customer order is associated with a unique parts requirements list $PRL = \{i_1, i_2, \ldots, i_N\}$, where N is the number of unique parts required to produce the vehicle, and $i_{l \in \{1,\ldots,N\}}$ denotes the quantity of each part. For example, $PRL = \{12, 3, 5\}$ means that part 1 is required twelve times, part 2 three times, and part 3 five times.

We use an accuracy measure to quantify the quality of our prediction of required parts as follows. First, we subtract the quantity of each part in the predicted PRL from the respective quantity of that part in the benchmark PRL. Then, we aggregate the absolute differences in quantities and divide the resulting value by the total quantity of parts occurring in the benchmark PRL. Let $PRL_{Benchmark} = \{I_1, I_2, \ldots, I_K\}$ and $PRL_{Prediction} = \{J_1, J_2, \ldots, J_K\}$ denote the benchmark PRL and the predicted PRL,

respectively, where K is the total number of unique parts in the union of both lists. Then, the difference between $PRL_{Benchmark}$ and $PRL_{Prediction}$ is given by

$$D := PRL_{Benchmark} - PRL_{Prediction} = \sum_{k=1}^{K} |I_k - J_k| \tag{6}$$

Given this notation, the accuracy A is

$$A = 1 - \frac{D}{\sum_{k=1}^{K} I_k} \tag{7}$$

Accuracy A gives the percentage of correctly predicted quantities within the benchmark PRL. For instance, if $A = 0.97$, the predicted PRL deviates by 3% from the benchmark PRL in the quantities of parts.

4 Evaluation

This section reports an experimental evaluation of the proposed data mining technique. We describe the setup, report the empirical results, and discuss the findings.

4.1 Experimental Setup

Our experiments used a unique data set of 47,499 actual orders received by a car manufacturer within a given time period. These orders contain fully customized car orders (i.e., including all configured options), associated with the specific PRL for each vehicle. For instance, in a random group of 20 orders, 9 vehicles were ordered with rear-view camera, 11 with active parking assist, 10 with cruise control, 2 with traffic sign recognition, and so on. The complete data set contained 55 different options for customers to choose from. Figure 1 shows an excerpt of the frequency of the configured options in this group taken from our unique data set.

We consider a scenario where the manufacturer does not know the exact option configurations for future orders. For example, the manufacturer estimates that from within 20 future orders, 5 orders contain a rear-view camera, 10 an active parking assist, and 12 a cruise control. At this point in time, the manufacturer does not have more information concerning all other potential options. In the presence of incomplete information, the manufacturer now wants to predict the quantities of those parts with long lead times that are required to produce these 20 vehicles.

To predict the PRL in this scenario, we randomly selected groups of varying sizes from {20, 100, 200, 500} as benchmark groups. Then, we systematically removed varying sets of options from these groups. By removing these sets of options, we mimic the incomplete option estimate provided by the sales manager. Next, we applied our data mining technique to identify the group in the historic order data set that is most similar to the benchmark group. Finally, we compared the aggregated PRL of the most similar group to that of the benchmark group. Figure 2 depicts the flow chart of the proposed data mining technique for predicting the PRL.

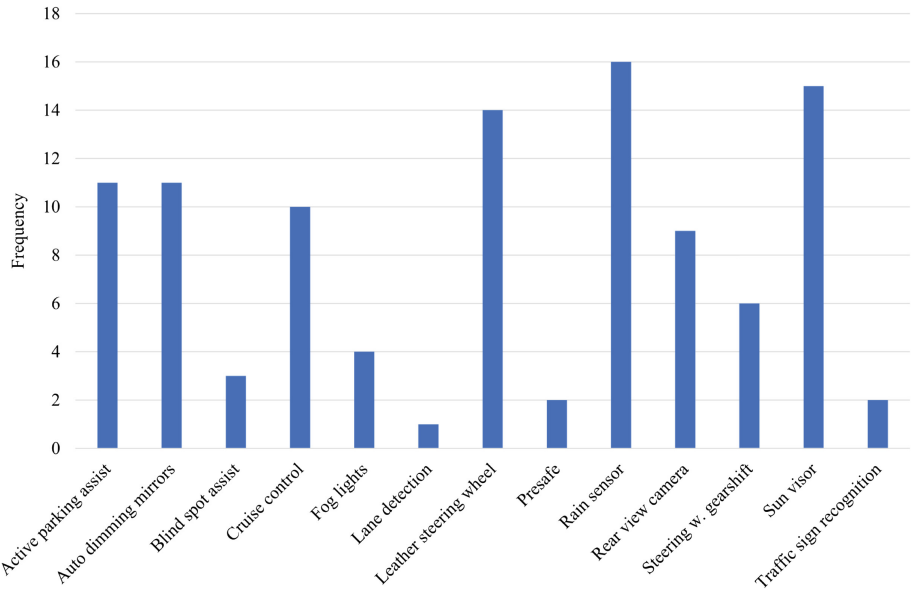

Fig. 1. Frequency of configured options in a random group of 20 orders (excerpt).

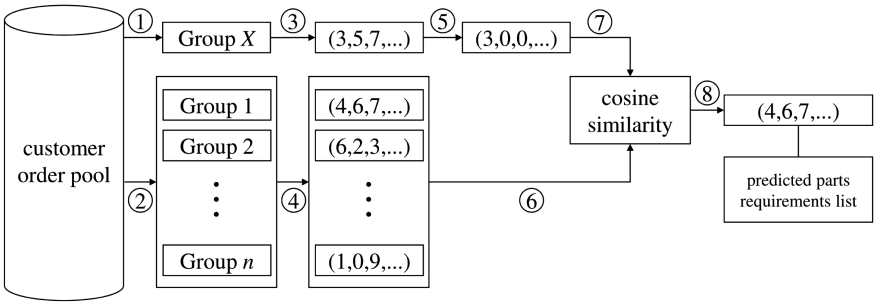

Fig. 2. Flow chart of the proposed data mining technique.

In step 1, we randomly select group X as the benchmark group, for which the PRL is known. Then, the historic order set is randomly divided into n distinct groups of equal size (step 2). Group X and groups 1 to n are then vectorized using the frequency of each option in the associated group (steps 3 and 4). In step 5, we define three uncertainty levels of the ordered vehicle configuration: low, medium, and high. For this purpose, we randomly remove a varying number of options from group X by setting the associated frequency to zero. When 13 out of 55 total options are removed, the level of uncertainty is said to be low (i.e., 23.6%). When 26 of 55 options are removed, the level of uncertainty is said to be medium (i.e., 47.3%). Finally, when 39 of 55 options are removed, the level of uncertainty is said to be high (i.e., 70.9%). In steps 6 and 7,

we determine the cosine similarity between group X and groups 1 to n to identify the group that is most similar to the "stripped" group X. Once the most similar group has been found, we compare the associated PRL to the PRL of the original group X to estimate the accuracy of the predicted part requirements.

4.2 Results

To validate our data mining technique, we calculated the accuracy of our prediction (of required parts) as a function of the associated cosine similarity. We divided the 47,499 orders into 475 groups of group size 100. One group was randomly selected as the benchmark group. Then, we calculated all angles between the benchmark group and the remaining groups. Next, we determined the accuracy of each group's PRL compared to the original PRL of the benchmark group. Figure 3 depicts the accuracy obtained for all 475 groups. Each point in the diagram corresponds to the accuracy obtained for a single group. The red line illustrates the trend line based on linear regression. As shown in Fig. 3, the accuracy of the prediction increases as the cosine similarity increases. Notice that increasing cosine similarities result in decreasing angles between vectors. This result implies that our data mining technique is valid and can be applied to the problem studied in this work.

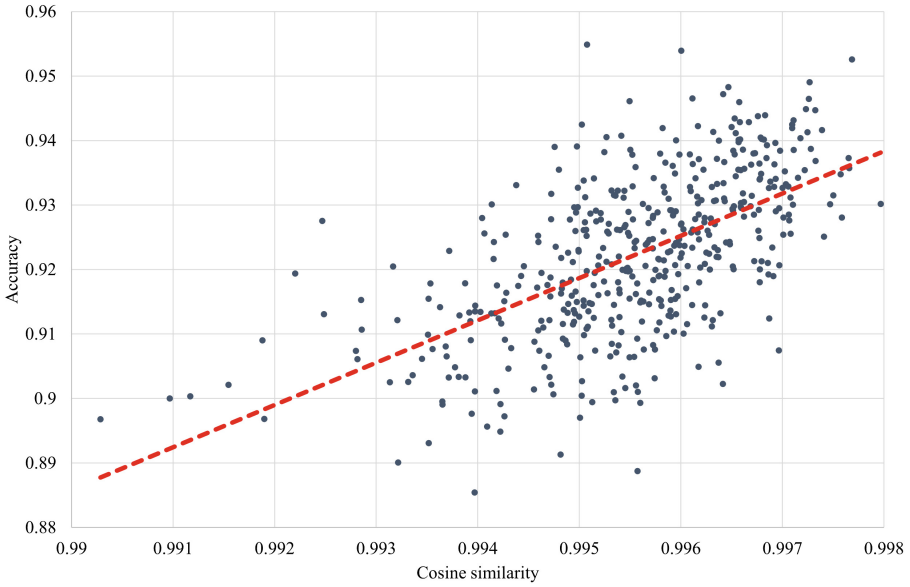

Fig. 3. Accuracy as a function of cosine similarity and linear regression trend line (red) (Color figure online).

After having successfully validated our approach, we now report on the results obtained in our simulation study. Table 1 presents the results of the simulation. For

each group size 20, 100, 200, and 500, we calculated the accuracy of the PRL for the scenarios “none”, “low”, “medium”, and “high.” Here, scenario “none” corresponds to a benchmark group that contains all options (no uncertainty). Thus, if that exact group were contained in the data set, the accuracy would be 100%.

Table 1. Accuracy for varying group sizes and uncertainty levels.

Group size	Accuracy for uncertainty level			
	0% (none)	23.6% (low)	47.3% (medium)	70.9% (high)
20	91.98	88.16	88.14	88.91
100	93.02	92.92	91.60	91.27
200	97.30	96.50	96.18	96.18
500	97.73	97.41	97.02	97.02

Figure 4 depicts the results graphically. For group size 20 (blue line), the accuracy is decreasing for increasing uncertainty level, reaching 88.14% at uncertainty level medium. Then, the accuracy increases up to 88.91% for uncertainty level high. When a group contains 100 orders (yellow line), the accuracy decreases for all uncertainty levels with its highest value of 93.02% down to its lowest value of 91.27%. For group size 200 (green line), the accuracy decreases from 97.3% to 96.18% for all uncertainty levels. Finally, when 500 orders are grouped (grey line), the accuracy also decreases for all uncertainty levels, falling from 97.73% to 97.02%.

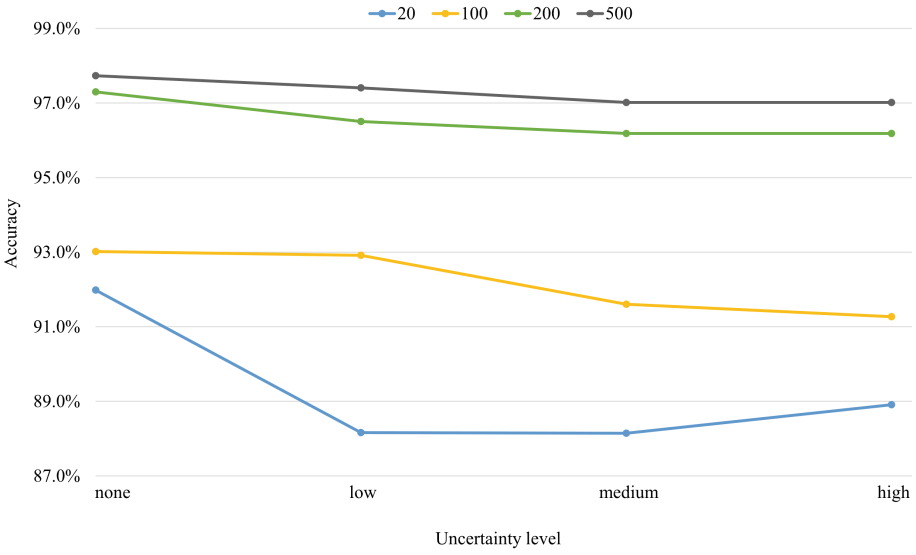

Fig. 4. Accuracy for different uncertainty levels at varying group sizes (Color figure online).

4.3 Discussion

Our experiments demonstrate the impact of uncertainty about future customer orders on the accuracy of predicting the material requirements for production in automotive industry applications. Our findings provide evidence for the efficacy of the proposed data mining technique to predict the quantities of parts with long lead times based on a large data set of historical customer orders. In the following paragraphs, we discuss the insights that can be obtained from our research.

First, we find that the proposed cosine similarity measure suits well for predicting material requirements. As the cosine similarity of vectors increases, the accuracy increases also (see Fig. 3). This finding implies that the frequencies of options within a historical customer order group correlate with the future requirements of parts subject to long lead times. The fact that the smallest observed cosine similarity between vectors is approximately 0.99 indicates that different order groups exhibit similar requirement lists for long lead time parts. In consequence, the quantities required to produce these vehicles can be predicted with an accuracy of close to 96%. This result indicates that potential violations of the assumption of linearly independent dimensions in the underlying vector space are not too detrimental to the accuracy results achieved by our technique.

Second, we find that larger customer order groups entail higher accuracy. As more customer orders are pooled, the quantities of parts and components required to produce these cars converge to those quantities in the associated benchmark group. This finding implies that the quantity of parts required for production becomes invariant as group sizes increase. In other words, individual parts and components are re-used by manufacturers for producing different car variants. This finding is consistent with the trend towards stringent modularization in the automotive production industry [10]. Car manufacturers implement modularization strategies to manage the increasing complexity and variant diversity of their vehicles by standardizing interfaces and individual components. As such, the requirements for mass customization can be achieved more effectively [25]. Moreover, smart production plants benefit from pervasive digital technology platforms as the central focus of the firm's innovation process. Car manufacturers can now use the same digital tools to design and control multiple modules and components that were dispersed among supplies in the past [11]. Hence, digitalized production promotes the development of innovative modularization concepts which in turn influences the prediction and procurement of material at distributed production plants in the automotive industry. The accuracy values obtained for increasing group sizes in our study thus help explain to what extent product modularity impacts the prediction of material requirements in automotive production.

Third, one advantage of our approach is that it can deal with high levels of uncertainty about the demand of possible option configurations. We find that accuracy decreases for increasing uncertainty levels. It is interesting to observe that for bigger group sizes the uncertainty level barely impacts the accuracy of the prediction. That is, if many customer orders are pooled, the number of parts and components required to produce this group of vehicles virtually matches that of the associated benchmark group. This finding corroborates that car manufacturers pursue a sustainable platform strategy for managing the complexity of their product variants. To this end, the

transformation toward digitalized production encourages the integration of data-driven analytics into business information systems to advance current prediction methods in automotive industry applications.

5 Conclusion

The contribution of this research is a data mining technique for predicting the requirements of parts with long lead times in the automotive industry. To evaluate our approach, we used a unique data set containing actual customer orders received by an international car manufacturer. In a first step, our approach incorporates the concept of cosine similarity to discover similar customer order groups within the data set. Then, we aggregate the quantities of the required parts and components for producing the vehicles within these different groups. Finally, we calculate the accuracy of our prediction relative to a predefined benchmark group. We find that increasing group sizes result in higher accuracy across all uncertainty levels. As car manufacturers continue to optimize product modularization using digital platform technologies, standardized parts and components for producing cars facilitate an improved prediction of material requirements even in the presence of uncertainty concerning future customer orders.

From a managerial perspective, our study can support supply chain managers in making more informed decisions about choosing the appropriate customer group size for predicting the demand in parts and components with long lead times. Because larger group sizes imply higher accuracy, managers can pool heterogenous estimates about future customer orders based on production capacity. At the same time, managers can focus on small sets of equipment options when forecasting material requirements because varying uncertainty levels show little impact on accuracy.

Future research can be pursued in four directions. First, our experimental evaluation could be extended by implementing *k*-means clustering on the data set [9, 22]. Unlike the random group formation used in our technique, the *k*-means clustering algorithm divides the data set into *k* clusters relative to the nearest mean. For benchmarking purposes, the accuracy obtained by *k*-means clustering could then be compared to the accuracy achieved in our study. Second, while the cosine similarity measure suits well to obtain high accuracy, other measures of similarity such as *k*-median or *k*-means + + algorithms could be used for comparing customer order groups. Third, for comparing parts requirements lists, different weights could be placed at different quantities of components. This change could provide further insights of how economic factors such as price and economies of scale affect parts requirement predictions. Fourth, future research can assess the applicability of our approach to other industries in which large quantities of parts are needed and customers can individualize the products.

Acknowledgements. This work has been partially supported by the Federal Ministry of Economic Affairs and Energy under grant ZF4541001ED8. We would like to thank Hansjörg Tutsch for his valuable comments on earlier versions of this paper. We also thank Lyubomir Kirilov for helping to develop and improve parts of this paper.

References

1. Lee, J., Lapira, E., Yang, S., Kao, A.: Recent advances and trends in predictive manufacturing systems in big data environment. Manuf. Lett. **1**(1), 38–41 (2013)
2. Ghabri, R., Hirmer, P., Mitschang, B.: A hybrid approach to implement data driven optimization into production environments. In: Abramowicz, W., Paschke, A. (eds.) BIS 2018. LNBIP, vol. 320, pp. 3–14. Springer, Cham (2018)
3. Renzi, C., Leali, F., Cavazzuti, M., Andrisano, A.O.: A review on artificial intelligence applications to the optimal design of dedicated and reconfigurable manufacturing systems. Int. J. Adv. Manuf. Technol. **72**, 403–418 (2014)
4. Dremel, C., Herterich, M.M., Wulf, J., Waizmann, J.-C., Brenner, W.: How AUDI AG established big data analytics in its digital transformation. MIS Quart. Exec. **16**(2), 81–100 (2017)
5. Leukel, J., Jacob, A., Karaenke, P., Kirn, S., Klein, A.: Individualization of goods and services: towards a logistics knowledge infrastructure for agile supply chains. In: Proceedings of the AAAI Spring Symposium (2011)
6. Widmer, T., Premm, M., Kirn, S.: A formalization of multiagent organizations in business information systems. In: Abramowicz, W., Alt, R., Franczyk, B. (eds.) BIS 2016. LNBIP, vol. 255, pp. 265–276. Springer, Cham (2016)
7. Meyr, H.: Supply chain planning in the German automotive industry. OR Spectr. **26**, 447–470 (2004)
8. Lee, H.L.: Aligning supply chain strategies with product uncertainties. Calif. Manag. Rev. **44**(3), 105–119 (2002)
9. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, 2nd edn. ACM Press, New York (1999)
10. Takeishi, A., Fujimoto, T.: Modularization in the auto industry: interlinked multiple hierarchies of product, production and supplier systems. Int. J. Automot. Technol. Manage. **1** (4), 379–396 (2001)
11. Yoo, Y., Boland, R.J., Lyytinen, K., Majchrzak, A.: Organizing for innovation in the digitized World. Organ. Sci. **23**(5), 1398–1408 (2012)
12. Kurbel, K.E.: MRP: material requirements planning. In: Swamidass, P.M. (ed.) Enterprise Resource Planning and Supply Chain Management, pp. 19–60. Springer, Heidelberg (2013)
13. Gupta, A., Maranas, C.D.: Managing demand uncertainty in supply chain planning. Comput. Chem. Eng. **27**(8–9), 1219–1227 (2003)
14. Zorgdrager, M., Curran, R., Verhagen, W., Boesten, B., Water, C.: A predictive method for the estimation of material demand for aircraft non-routine maintenance. In: 20th ISPE International Conference on Concurrent Engineering (2013)
15. Lee, Y.Y., Kramer, B.A., Hwang, C.L.: Part-period balancing with uncertainty: a fuzzy sets theory approach. Int. J. Prod. Res. **28**(10), 1771–1778 (1990)
16. Du Chih-Ting, T., Wolfe, P.M.: Building an active material requirements planning system. Int. J. Prod. Res. **38**(2), 241–252 (2000)
17. Steuer, D., Korevaar, P., Hutterer, V., Fromm, H.: A similarity-based approach for the all-time demand prediction of new automotive spare parts. In: 51st Hawaii International Conference on System Sciences (HICSS 2018), pp. 1525–1532. Waikoloa Village (2018)
18. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
19. Manning, C.D., Ragahvan, P., Schutze, H.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009). Online ed

20. McCallum, A., Nigam, K.: A comparison of event models for naive Bayes text classification. In: 15th National Conference on Artificial Intelligence (AAAI 1998): Workshop on Learning for Text Categorization, pp. 41–48, Madison (1998)
21. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS (LNAI), vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
22. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
23. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Annals Data Sci. **2**(2), 165–193 (2015)
24. Riekert, M., Leukel, J., Klein, A.: Online media sentiment: understanding machine learning-based classifiers. In: 24th European Conference on Information Systems (ECIS 2016), Istanbul (2016)
25. Tu, Q., Vonderembse, M.A., Ragu-Nathan, T.S., Ragu-Nathan, B.: Measuring modularity-based manufacturing practices and their impact on mass customization capability: a customer-driven perspective. Decis. Sci. **35**(2), 147–168 (2004)