Finding Demand Patterns in Supply Chain Planning – Two Clustering Approaches for Semiconductor Production Industry

Article · August 2018

DOI: 10.17560/atp.v60i09.2360

5 authors, including:

Sebastian Achter
Technische Universität Hamburg

Thomas Ponsignon
Infineon Technologies

Hans Ehm
Infineon Technologies

Matthias Meyer
Technische Universität Hamburg

Comparison of two clustering approaches to find demand patterns in semiconductor supply chain planning

Pramod Govindaraju, Sebastian Achter, Thomas Ponsignon, Hans Ehm, and Matthias Meyer

***Abstract*— Advancements in semiconductor industry have resulted in the need for extracting vital information from vast amount of data. In the operational process of supply chain, understanding customer demand data provides important insights for demand planning. Clustering analysis offers potential to identify latent information from multitudinous customer demand data and supports enhanced decision-making. In this paper, two clustering algorithms to identify customer demand patterns are presented, namely K-means and DBSCAN. The implementation of both algorithms on the prepared data sets is discussed and their performance is compared. The paper outlines the importance of deciphering valuable insights from customer demand data in the betterment of a distributed cognitive process of demand planning.**

I. Introduction

Due to persistent advancements in storage technology and computational power, organizations across various industries face a surge in data availability. Data from almost all organizational processes, including marketing, material planning and control, production scheduling and maintenance, etc. are recorded. Generating and managing knowledge from such vast amounts of data can be a valuable asset for an enterprise in differentiating itself from competitors [1] and enhancing its own process efficiency.

The semiconductor industry is a highly competitive market. High demand volatility, rapidly changing environments as well as contracting product life cycles challenge the supply chain (SC) planning process [2]. Manufacturing lead-times (months) are usually longer than customer order lead-times [3]. Products and processes increase in intricacy with every cycle. These market properties amplified by globalization, diversity of variants, and declining manufacturing depth [4] result in a highly complex SC. Within such a complex environment, competitive pressure gives rise to a continuous endeavor for cost reduction which is promised to be primarily achieved through advancements in the operational process [5].

The manufacturing process in the semiconductor industry is already characterized by a high degree of automation. Hence, different approaches of data mining are already applied and proofed their viability e.g. in reducing power consumption [6] or achieving quality improvements [7] [8] [9]. However, also the operational processes benefit from applications of data mining techniques, e.g. for optimizing resource allocation [10] or work-in-progress inventory levels [11].

We identify the operational demand planning process as an area in which human judgment and decision making induces the flexibility that is essential for ensuring stability and innovation in an uncertain and complex environment [12]. We specifically look at the role of the Operational Demand Planner (ODP). The ODP is responsible for the demand planning by entering demand forecasts and stock targets that are used later in the Advanced Planning and Scheduling System (APS). To do so, the ODP uses information about the customer demand from different organizational reporting and interacts with other decision makers in his environment. The ODPs decisions influence the planning performance both negatively, e.g. due to biases or improper personal strategies as well as positively, e.g. by drawing on informal domain knowledge or flexible behavior in unusual situations [12]. Hence, the generated data entered in the APS originate from behavioral responses of ODP as a human decision maker embedded in the environment of the demand planning process. An illustration of the distributed cognitive process of ODP in demand planning is shown in Figure 1.

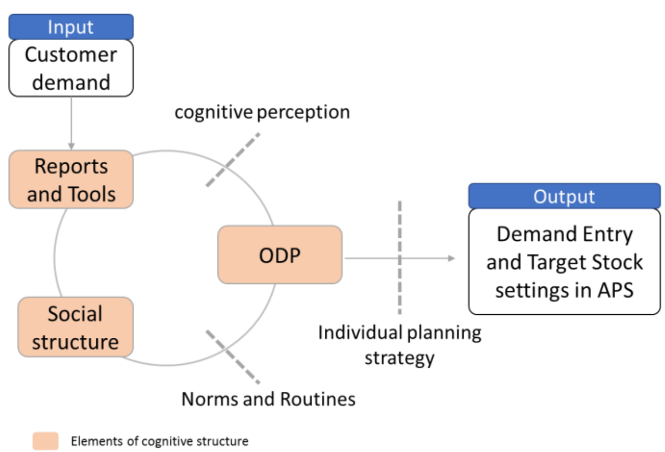


Figure 1. Distributed cognitive process of operational demand planner

Although the ODP has a key role in the planning process, crucial behavioral or cognitive traits could not yet be clearly identified. The cognitive process is unobservable and social norms or routines are hard to unveil. Whilst an extensive qualitative study might offer one approach to discover behavioral patterns, we try to find such patterns implicitly in the data beforehand. This approach is inspired by Hutchins' distributed cognition theory, which assumes that cognitive structures are distributed over social (e.g. interaction among people), physical (e.g. tools and reporting) and timely (iterative adjustment of the structure) dimensions [13].

Pramod Govindaraju, Thomas Ponsignon, and Hans Ehm are with Infineon Technologies AG, Neubiberg, 85579 Germany (corresponding author: Thomas Ponsignon; phone: +4989234-27371; fax: +4989234-9552300; e-mail: Thomas.Ponsignon@infineon.com).

Sebastian Achter and Matthias Meyer are with the Hamburg University of Technology, Hamburg, 21073 Germany.

We therefore aim to find categories of demand patterns based on the customer demand data, which represent the main input for the planning process. In the next step, we look for dominant planning strategies in the demand entry and target stock settings by ODP in APS within the identified customer demand groups. This approach allows us in the identification of dominant planning strategies while treating unobservable data of the cognitive process as a black box. Such findings can be used to give direction for structured qualitative studies or to support ODPs in future decision making and pave the way for increasing automation using supervised learning techniques such as neural networks.

The goal of the paper is to introduce our approach for deriving product groups from customer demand pattern, using different clustering techniques, as a first step towards the identification of dominant planning strategies of the ODP. Clustering, which is an unsupervised learning technique, is a viable option in our case because we do not have prior information on the number of product groups and the class labels of the data [14]. Since there is a variety of possible algorithm available we compare the results of preselected algorithms, namely K-means and DBSCAN, which we believe are suitable for the researched problem.

II. Clustering Process

Based on the initial framework on Knowledge Discovery in Databases [15], the methodical steps involved in cluster analysis has been defined in the literature [16]. An illustration of the steps in clustering process is shown in Figure 2.

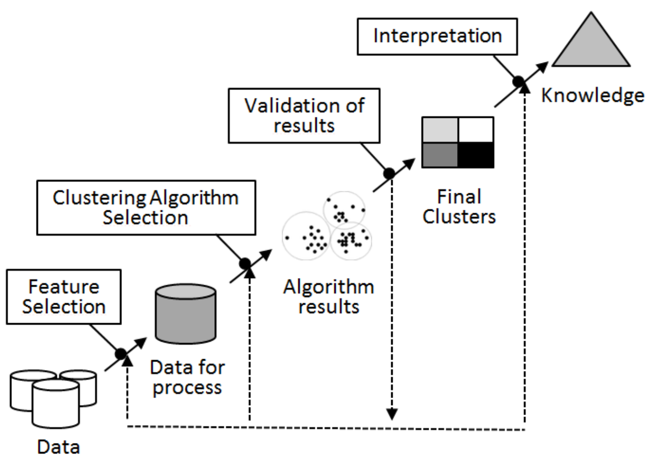


Figure 2. Steps of clustering process [16]

Clustering process is mainly composed of the following four steps:

1. Feature selection: The step refers to the selection of features to perform the cluster analysis. In cluster analysis, since the class labels are not predefined, there are high chances of selecting redundant and irrelevant features. Furthermore, elimination of unnecessary features enhances clustering results [17].

2. Clustering algorithm: The choice of a clustering algorithm influences the clusters obtained from the data [14]. The result obtained from clustering algorithms are based on some assumptions depending on the properties of the data set (geometry and density distribution) and input parameter values [18] since the class labels are not specified.

3. Validation of the results: In this step, the clusters obtained from the algorithm are evaluated. Generally, visualizing the clusters serves as a good step to quickly verify the cluster results. However, for large multidimensional data, effective visualization becomes arduous.

   Validity indices are commonly used for measuring the quality of cluster results [19]. There are two types of cluster validation indices: external indices and internal indices. External indices require a prior knowledge of the cluster structure in the data whereas the internal indices do not require a prior knowledge of the data [20]. Due to this reason, the optimal number of clusters is usually determined from the internal indices.

4. Interpretation of the results: It has been highlighted in the literature [16] [14] that in order to draw better conclusions from the validated clusters, domain knowledge of the experts serves as an invaluable input. Furthermore, interpretation could also be extended by integrating the expert reviews with other experimental evidence and analysis.

III. Cluster Analysis On The Example Of Infineon

Customer demand data, which is mainly obtained from reporting, serve as an important source of information for the ODP in demand planning process. Different measures can be obtained from the customer demand data, namely (i) Order quantity, (ii) Order frequency, (iii) Order lead time (OLT), and (iv) Customer Forecast Accuracy. The forecast accuracy of customer is a derived measure based on the customer forecast and customer order which is calculated in our case using SMAPE3 [3]. Demand pattern is a composition of the aforementioned measures of the customer demand.

Cluster analysis has been adapted at Infineon to segregate similar product groups based on characteristics of customer demand data. As a first step in implementing the cluster results, the planning performance of ODPs for the different product groups was studied to understand the influence of demand patterns on ODP planning. The outcome of the study was that the ODPs were not able to control the planning performance of product groups with unstable demand. Primarily, the performance measurement of ODP is dependent on several other activities in addition to the demand planning process. In order to focus only on the demand planning process, the objective now is to use cluster analysis in understanding the influence of demand pattern on ODP planning process. The study helps in determining whether ODPs input for demand entry and target stock settings in APS is dependent on the clustered product groups with different demand patterns.

In the first study on ODP planning performance [21], two commonly used algorithms namely Average Linkage and K-means were used to cluster the product groups. Based on the internal validation method using Calinski-Harabasz index, the optimum number of clusters for K-means and Average Linkage algorithms was determined. The evaluation of the clusters based on the reviews of ODP showed that K-means performed better than the Average Linkage algorithm. Although K-means has high computational efficiency, low memory consumption and ease of implementation [22], the algorithm is not robust against outliers and is capable of detecting only convex shaped clusters. As a result, another advanced algorithm DBSCAN is considered for clustering. DBSCAN is robust against outliers, capable of detecting arbitrary shaped clusters and has good efficiency for large spatial database [23] [24]. Furthermore, there is no requirement to know the number of clusters in the data *a priori*, as opposed to K-means [25]. In the ongoing study on ODP planning process, cluster analysis is performed using both K-means and DBSCAN algorithms to understand if the recently introduced DBSCAN algorithm performs better than the traditional K-means algorithm.

Clustering Methodology

The methodology adopted for cluster analysis at Infineon is based on the steps illustrated in Figure 2. The software used for the cluster analysis was R-Studio. The description of the steps is provided below:

1. Feature selection: The features required for the cluster analysis is derived from the information defining the demand pattern. In order to exclude the redundant features, the degree of association between a pair of features was measured using Spearman correlation coefficient. In case of feature pairs with high correlation, one of the features was excluded to remove redundancy. In addition, an alternative approach for feature reduction, principal component analysis (PCA) [26] was also implemented. However, the approach was not considered since the clusters cannot be interpreted easily using the features obtained from PCA. The selected features are; Average Billings, Coefficient of Variation (CoV) of Average Billings, Frequency of Billings, Maximum OLT, Minimum OLT, Forecast accuracy and CoV of Forecast Accuracy.
2. Data preparation: Based on the selected features, two data sets are obtained for cluster analysis. Firstly, the features in the data sets are scaled to prevent the discrepancies in clustering results due to their varying magnitudes. In the next step, the outliers from the data sets are separated because it is understood that K-means is not robust against outliers. Applying K-means algorithm on a data set with outliers does not reflect the true potential of the algorithm. The outliers in data sets are identified using Inter-Quartile Range (IQR) rule. According to this rule, the data points below Quartile 1-1.5*IQR and above Quartile 3 + 1.5*IQR are considered as outliers. In total, the two data sets are divided into four data sets after separating the outliers.
3. Cluster analysis: K-means and DBSCAN algorithms are applied for the analysis. The input parameter for K-means algorithm is the number of clusters ‘k’. Internal validity approach is used to determine the optimum value of k from Calinski-Harabasz (CH) index. CH Index value for the four data sets is calculated for the range of k between 2 to 10. The value of k corresponding to highest CH index is considered as the input parameter. The two input parameters for DBSCAN are minimum number of points (μ) and radius (ε). The value of μ is set to 2 because a lower μ value facilitates clustering of outliers as well. The ε value is determined based on μ and k-nearest neighbor algorithm [27]. The idea of this approach is to calculate the average of the distances of every point to its k-nearest neighbors where the value of k corresponds to μ. As a next step, these k-distances are plotted in an ascending order. The aim is to determine the “knee”, which corresponds to the optimal radius (ε).

TABLE I. INPUT PARAMETERS FOR K-MEANS AND DBSCAN

Data set	Datapoints	K-means	DBSCAN
1	No outliers	k=2	μ=2, ε=1.1
	Outliers	k=8	μ=2, ε=2.6
2	No outliers	k=2	μ=2, ε=1.4
	Outliers	k=6	μ=2, ε=2.4

4. Cluster evaluation: Visualization serves as a good first step in understanding the cluster results. However, the evaluation of the clusters obtained from the two algorithms is based on the review obtained from ODP.
5. Interpretation: After evaluating the clustered product groups obtained from both algorithms, the characteristics of the clusters are determined using the features selected. The association of clusters with the features enables interpreting the characteristics of demand patterns associated to the product groups.

IV. RESULTS

A comparison of the results obtained from the algorithms K-means and DBSCAN indicated that the clusters obtained from DBSCAN were better than those of K-means. The validation of the results was based on the review obtained from ODP. It was further noticed that DBSCAN provided significantly better results for the data set with outliers in comparison to K-means. As seen in Figure 3, the data points in a cluster from K-means algorithm are widespread indicating lesser compactness. Furthermore, the inherent nature of the algorithm to allocate all the data points to one of the clusters hampers the cluster results.

The clusters from DBSCAN as shown in Figure 4 are well separated and more compact. Moreover, the data points not belonging to any of the clusters are shown as black dots. These data points can be considered as distinct products with different characteristics. The significance of this result is that these extreme points are not blended in different clusters and can be easily identified.

In Figures 3 and 4, the x- and y-axes are denoted by Dim 1 and Dim 2, respectively. The axes are suggested by the software to depict maximum information for cluster visualization and do not represent any of the features selected. The reason for this depiction is that clustering process is carried on a multidimensional data and it is not possible to visualize the results with all the dimensions in a single plot.

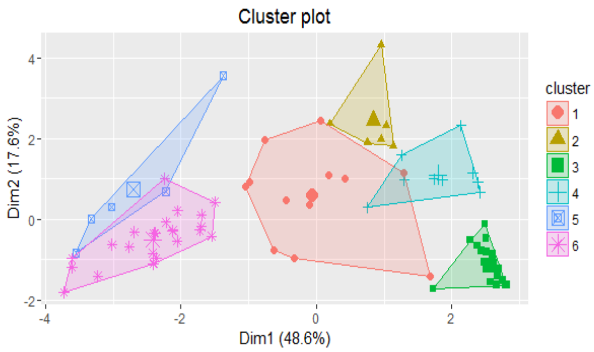
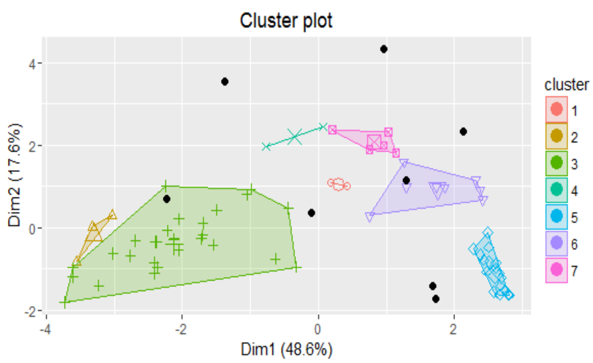


Figure 3. K-means clusters for outliers in Data set 2

Figure 4. DBSCAN clusters for outliers in Data set 2

V. Conclusion



In summary, we have obtained results indicating better performance of DBSCAN over K-means in clustering the product groups based on the demand patterns.

The next step of the study is to find the pattern of demand entry and target stock settings for the clusters obtained from DBSCAN algorithm. In doing so, the influence of demand pattern on the demand planning process can be understood and the important features characterizing the product groups can be identified. This information would serve as an important input in defining the rules based on the features for demand planning process. Furthermore, the method can be used as an input for supervised learning techniques such as neural networks and vector component analysis thereby increasing the scope for automation. In the future course of the study, we also plan to also integrate soft clustering algorithms such as Fuzzy C-Means algorithm to identify the demand patterns.

References

[1] J. A. Harding, M. Shahbaz, Srinivas, and A. Kusiak, “Data Mining in Manufacturing: A Review,” *J. Manuf. Sci. Eng.*, vol. 128, no. 4, pp. 969–976, Dec. 2005.

[2] N. Geng and Z. Jiang, “A review on strategic capacity planning for the semiconductor manufacturing industry,” *Int. J. Prod. Res.*, vol. 47, no. 13, pp. 3639–3655, Jul. 2009.

[3] H. C. Ott, S. Heilmayer, and C. S. Y. Sng, “Granularity dependency of forecast accuracy in semiconductor industry,” *Res. Logist. Prod.*, vol. 3, 2013.

[4] E. D. Beinhocker, *Managing in a complex world*. The Mckinsey Quarterly, 2007.

[5] L. Mönch, M. Stehli, J. Zimmermann, and I. Habenicht, “The FABMAS multi-agent-system prototype for production control of water fabs: design, implementation and performance assessment,” *Prod. Plan. Control*, vol. 17, no. 7, pp. 701–716, Oct. 2006.

[6] C. M. Yu, C. F. Chien, and C. J. Kuo, “Exploit the Value of Production Data to Discover Opportunities for Saving Power Consumption of Production Tools,” *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 4, pp. 345–350, Nov. 2017.

[7] G. Köksal, \.Inci Batmaz, and M. C. Testik, “Review: A Review of Data Mining Applications for Quality Improvement in Manufacturing Industry,” *Expert Syst Appl*, vol. 38, no. 10, pp. 13448–13467, Sep. 2011.

[8] C.-F. Chien, W.-C. Wang, and J.-C. Cheng, “Data Mining for Yield Enhancement in Semiconductor Manufacturing and an Empirical Study,” *Expert Syst Appl*, vol. 33, no. 1, pp. 192–198, Jul. 2007.

[9] C.-F. Chien, C.-W. Liu, and S.-C. Chuang, “Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement,” *Int. J. Prod. Res.*, vol. 55, no. 17, pp. 5095–5107, Sep. 2017.

[10] C. M. Yu and C. J. Kuo, “Data mining approaches to optimize the allocation of production resources in semiconductor wafer fabrication,” in *2016 International Symposium on Semiconductor Manufacturing (ISSM)*, 2016, pp. 1–4.

[11] Y.-H. Lin, J.-R. Shie, and C.-H. Tsai, “Using an Artificial Neural Network Prediction Model to Optimize Work-in-process Inventory Level for Wafer Fabrication,” *Expert Syst Appl*, vol. 36, no. 2, pp. 3421–3427, Mar. 2009.

[12] S. Achter *et al.*, “On agent-based modeling in semiconductor supply chain planning,” in *2017 Winter Simulation Conference (WSC)*, 2017, pp. 3507–3518.

[13] E. Hutchins, *Cognition in the Wild*, vol. 7. MIT Press, 1995.

[14] S. Theodoridis and K. Koutroumbas, *PATTERN RECOGNITION*, 2nd ed. United States of America: Elsevier, 2003.

[15] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework.,” in *KDD*, 1996, vol. 96, pp. 82–88.

[16] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *J. Intell. Inf. Syst.*, vol. 17, no. 2–3, pp. 107–145, 2001.

[17] J. G. Dy and C. E. Brodley, “Feature selection for unsupervised learning,” *J. Mach. Learn. Res.*, vol. 5, no. Aug, pp. 845–889, 2004.

[18] M. Halkidi and M. Vazirgiannis, “A Data Set Oriented Approach for Clustering Algorithm Selection,” presented at the Principles of Data Mining and Knowledge.PKDD 2001 Discovery, 2001, vol. 2168, pp. 165–179.

[19] K. Wang, B. Wang, and L. Peng, “CVAP: validation for cluster analyses,” *Data Sci. J.*, vol. 8, pp. 88–93, 2009.

[20] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, “Internal versus external cluster validation indexes,” *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.

[21] S. Achter, “Influence of Demand Patterns on the Planning Performance in a Semiconductor Supply Chain,” M.S. thesis, Inst. of Mgmt. Ctrl. and Simul. , Hamburg University of Technology, Hamburg, 2017.

[22] L. Morissette and S. Chartier, “The k-means clustering technique: General considerations and implementation in Mathematica,” *Tutor. Quant. Methods Psychol.*, vol. 9, no. 1, pp. 15–24, 2013.

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters.” AAAI, 1996.

[24] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, “A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases,” *Int. J. Comput. Appl.*, vol. 3, no. 6, Jun. 2010.

[25] M. H. Zafar and M. Ilyas, “A Clustering Based Study of Classification Algorithms,” *Int. J. Database Theory Appl.*, vol. 8, no. 1, pp. 11–22, Feb. 2015.

[26] C. Ding and X. He, “K-means clustering via principal component analysis,” in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.

[27] S. Kurumalla and P. S. Rao, “K-NEAREST NEIGHBOR BASED DBSCAN CLUSTERING ALGORITHM FOR IMAGE SEGMENTATION,” *J. Theor. Appl. Inf. Technol.*, vol. 92, no. 2, p. 395, 2016.