EARLY DEFECT DETECTION USING CLUSTERING ALGORITHMS[1]

Blanka Bártová, Vladislav Bína*

Abstract

Product quality is a crucial issue for manufacturing companies, so it is essential to take note of any emerging product defects. In contrast to the use of traditional methods, the "modern" constantly evolving data mining methods are now being more frequently used. The main objective of this paper is to detect the potential cause or the area of the production process where the majority of product defects arise. The dataset from the semiconductor manufacturing process has been used for this purpose. First, it was necessary to address dataset quality. Significant multicollinearity was found in the data and to detect and delete the collinear variables, correlations and variance inflation factors have been used. The MICE-CART method has been used for the imputation because the original dataset contained more than 5% of random missing values. In further analysis, the K-means clustering method has been used to separate the failed products from the flawless ones. Following this, the hierarchical clustering method has been used for the failed product to create groups of product defects with similar properties. For the optimal number of clusters, the determination of the BIC method has been used. Five clusters of products have been made although only three can be classed as important for further analysis. These groups of products should be directly subjected to the analysis in the production process, which can assist in identifying the source of scarcity.

Keywords: manufacturing, data mining, clustering, product quality, quality management, MICE-CART, VIF

JEL Classification: C38, C44, D24, L15

Introduction

The last decades have seen life undergoing a turbulent and fast-changing environment. Nowadays, due to rapid technological changes, automation, and robotics, a new technological revolution is taking place. A new era of the phenomenon of Industry 4.0 and smart factories is in progress and companies now face many challenges, such as short product life cycles, volatile demand and high customisation (Gaub, 2016). High-value manufacturing processes are increasingly moving towards flexible, intelligent production systems. To compete in future markets, manufacturing companies should be

1 This research was supported by an internal grant of the University of Economics in Prague IG32029 [F6/2/2019].

* University of Economics, Prague, Faculty of Management (blanka.bartova@vse.cz; vladislav.bina@vse.cz).

able to produce small batch sizes of a product or even a single item in a timely and cost-effective manner. They need to have sufficient functionality, scalability, and connectivity with customers and suppliers to meet these requirements (Schumacher et al., 2016). At the same time, to stay or strengthen the position of an organisation on the market, a modern business needs to follow the principles of quality control in its actions. In addition, to meet such challenges, systems will become more complex and difficult to monitor and control (Mabkhot et al., 2018). It is now common manufacturing practice to reduce and minimise the number of defects and errors in a process and to do things precisely at the first attempt. The ultimate aim is to reduce the number of defective products (Wang, 2013).

In this research, we focus on defects detected in manufacturing companies, specifically in companies in the metalworking industry. This paper proposes a data mining-based knowledge discovery approach using a sequence of two different types of clustering methods for detecting the major groups of products with a similar cause.

1. Literature Review

Quality is a term that is complex and difficult to specify. The word quality has many meanings, such as a degree of excellence, conformance with requirements, the totality of the characteristics of an entity that impact its ability to satisfy stated or implied needs, fitness for use, freedom from defects, imperfections or contamination and delighting customers (Hoyle, 1994). Various authors explain this notion differently. One of the "gurus" in quality control, William Edwards Deming (1982), defines quality as a predictable degree of uniformity and dependability at low cost and suited to the market. According to the American Society for Quality and Goetsch and Davis (2010), quality denotes excellence in goods and services, especially to the degree that they conform to the requirements and satisfy customers. The definition of quality stated by the International Organisation for Standardization (ISO) is: "The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (AS/NZS ISO, 1994, p. 7). Put more simply, one can say that a product has good quality when it complies with the requirements specified by the client (Knowles, 2011). In this research, we view quality as the compliance of product properties and dimensions with pre-specified company standards. Product quality is a crucial issue for manufacturing companies. It is essential for customer satisfaction, and so is directly connected with the company's revenues and market share. Quality is also closely connected to company performance. Sadikoglu and Zehir (2010) stated that Quality Management (QM) is a systematic, proven approach to improvements in organisational performance. Numerous empirical studies have attempted to investigate the relationship between QM practices and company performance (Mehran and Mehran, 2013).

There are many traditional approaches to quality management although nowadays data mining methods have become more useful and successful in manufacturing companies. Traditional methods, such as Total Quality Management (TQM) focuses on quality for customer satisfaction and concurrently sustains a company's competitive advantage in today's challenging and dynamic business environment (Yin et al., 2018). Other methodology such as Lean Six Sigma combines the Six Sigma techniques, which enable companies to reduce manufacturing defects with the lean manufacturing principles to help companies benefit from faster processing for lower costs with superior quality

(Dragulanescu and Popescu, 2012). However, even traditional "Six-Sigma" approaches cannot eliminate all the defects in manufacturing, only a very small share, given their limitation in dealing with complex and dynamic datasets. The Zero Defects concept developed by Philip Crosby (1979) means flawless production. The concept consists of preventing the occurrence of defects and flaws in all production stages. Quality management tools must be used to achieve this (Wang, 2013). All these methods are powerful tools for quality improvement but now at the time of Industry 4.0 and smart factories, which already have highly elaborate quality control process, these methods are no longer appropriate to use. In smart factories, mass data collected from various sensors and critical manufacturing-related knowledge can be hidden in the data. An example of such knowledge can include rules or regulations for identifying defects to the quality of the products. Human operators may never find the rules through manual investigation. This means they may never discover such hidden knowledge from the data. Traditional data analysis methods are no longer the best alternative to be used (Wang et al., 2006).

Quality improvement (QI) of industrial products and processes requires collection and analyses of data to solve quality related manufacturing problems. Traditional statistical process control approaches are less effective than data mining, especially when dealing with multivariate and autocorrelated processes (Evans, 2015). With the continual increase in process complexity, this inefficiency is becoming more apparent. A special multivariate and autocorrelated process is a process occurring within a heterogeneous production environment (a variety of types of machines, pots, etc. used for the same task). This makes the quality control of such processes more difficult (Horvath and Vircikova, 2012). Although traditional data analysis tools have been successfully used to improve the quality of products and processes, better tools now exist to mine massive data sets collected through computerised systems in the industry (Köksal et al., 2011). Data mining tools can be highly beneficial for discovering interesting and useful patterns, even in complicated manufacturing processes. However, data accumulated in manufacturing plants has unique characteristics, such as an unbalanced distribution of the target attribute, and a small training set relative to the number of input features. Thus, conventional methods are inaccurate in quality improvement cases (Choudhary et al., 2009). Data mining tools are useful in many areas of manufacturing such as defect analysis, yield improvement, quality monitoring, and process control, etc. (Rokach et al., 2008). Data mining tools can be used to extract knowledge from process data sets. The knowledge acquired can be used to minimise the number of defective products and to achieve the desired level of process performance and product quality (Ramana and Reddy, 2012).

2. Data Mining Application for Defect Detection

There are many data mining methods that are useful for application in manufacturing. Rough set theory or clustering analyses are frequently used to solve defect detection problems in manufacturing neural networks, association rules, and types of regression. For example, the journal paper, proposed by Bhuvaneswari and Sabarathinam (2013), examines the detection of defects in manufactured ceramic tiles to ensure high-density quality. The problem is concerned with the automatic inspection of ceramic tiles using an Artificial Neural Network (ANN). A detailed comparison between traditional statistical methods, the RST approach, and the extended RST approach is presented by Tseng,

Jothishankar and Tong (2004). The developed algorithm was applied to an industrial case study involving quality control of printed circuit boards (PCB), especially solder ball defects. The paper written by Sabet et al. (2017) presents a method for identifying unknown patterns between the manufacturing process parameters and the defects of the output products. The proposed method of fuzzy association rules also identifies the relationships between the defects.

Many articles on the implementation of clustering methods for defect detection in manufacturing have already been published, such as *Defect Segmentation of Semiconductor Wafer Image Using k-Means Clustering* by Saad et al. (2015). The K-means clustering partitional method used to identify and classify bearing defects was examined by Yiakopoulos et al. (2011). Another example of clustering use is the condition monitoring architecture of dynamical systems with unknown gradual faults using a dynamical clustering algorithm, which allows a continuous update of the operating modes of the system, was proposed by Chammas et al. (2014). The use of a combination of clustering algorithms with another method also became more frequent. The new method using the K-means clustering algorithm in combination with a self-organising map was used by Saludes-Rodil (2015) for the classification of surface defects in wire rod production. Yusof et al. (2018) in his research applied the principle component analysis (PCA) as a pre-processing method for hierarchical clustering analysis on the frequency spectrum of the vibration signal.

According to the literature review, we conclude that application combinations of cluster analysis for defect detection problem solving are not as usual as the other data mining methods mentioned above or clustering analysis itself. The aim of this paper is to propose the combination of two different types of a clustering algorithm for the detection of poor-quality products and the creation of groups of products with a similar cause of errors. To achieve the aim, we have formulated the following research question: *Can we identify groups of poor-quality products with the same cause of error by using a sequence of two different clustering algorithms?*

3. Dataset

The dataset used in this paper is from a complex modern semiconductor manufacturing SECOM process (McCann and Johnston, 2008). These are records of the monitoring of signals/variables collected from sensors and process measurement points. However, not all these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information plus irrelevant information as well as noise. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then the feature selection can be applied to identify the most relevant signals. The process engineers can then use these signals to determine key factors contributing to yield excursions downstream in the process. The dataset presented in this case is a selection of those features where each example represents a single production entity with associated measured features.

There are 1567 examples taken from a wafer fabrication production line. There are both failed and passed products in the quality control system. For product quality, 590 measuring sensors and process measurement points (variables) were used. In other words, each example is a vector of 590 sensor measurements. This results in a dataset of 924530 values measured during the production process. For such a large volume of

measurement data, automatic fault detection technique is essential. The large amount of metrology data obtained from hundreds of sensors make this dataset difficult to accurately analyse. Thus, our main focus is to devise a method based on data mining techniques to build an accurate model for fault detection. There are also 5% of missing values and collinear variables in the data set, which is necessary to be resolved before the clustering.

Various papers using this dataset have been already published, such as *Feature Selection and Boosting Techniques to Improve Fault Detection Accuracy in the Semiconductor Manufacturing Process* written by Kerdprasop and Kerdprasop (2011). In this paper, the authors investigate the application of data mining techniques such as decision tree induction, naïve Bayes analysis, logistic regression, and k-nearest neighbour classification to create an accurate model for fault case detection. Further research using the same data set is *Quality prediction modelling for multistage manufacturing based on classification and association rule mining* written by Kao et al. (2017) in which the authors introduce a framework for quality prediction modelling in a multistage manufacturing system (MMS) environment.

For cluster analysis implementation, R studio software has been used. First, we will make the data cleaning, deleting irrelevant variables and imputing the missing values. Then, using the K-means clustering method, we will split the monitored products into two groups: one group of failed products and one group of flawless products. The cluster analysis will then be applied only to the dataset of failed products. We will apply the hierarchical clustering method and will change the settings in the method.

4. Methodological Approach

First, it is necessary to prepare the data set for the following analysis. For this purpose, the method of data imputation will be chosen. Then we will apply different types of clustering methods on the data and make a comparison. Several variants of algorithm settings will be used.

4.1 Preparing and Cleaning Data

The simplest solution for the missing values imputation problem is the reduction of the data set and the elimination of all missing values. This can be done by eliminating the samples (rows) with missing values (Kantardzic, 2003) or eliminating the attributes (columns) with missing values. Both approaches can be combined. Elimination of all samples is also known as complete case analysis (Kaiser, 2014). In this case, we will reduce the attributes because there are many constant attributes and collinear variables. We will use the Variance Influence Factor (VIF) method for multicollinearity reduction in the dataset (Paul, 2006). We can compute the VIF with the formula, where the symbol R_i^2 means the coefficient of determination and analyse the magnitude of multicollinearity by considering the size of VIF_i. A rule of thumb is that if $VIF_i > 5$ then multicollinearity is high (Kuther et al., 2005). Variables with a high VIF will be deleted.

$$VIF_i = \frac{1}{1 - R_i^2}. \tag{1}$$

After the reduction of the data set, we can proceed to the data imputation. There are two basic types of imputation: single and multiple. Single refers to a single estimate of the missing value and is popular because it is conceptually simple and because the resulting sample has the same number of observations as the full data set (D'ambrosio et al., 2012). Some imputation methods result in biased parameter estimates, such as means, correlations, and regression coefficients, unless the data is MCAR (Missing completely at random). The bias is often worse than with listwise deletion, the default in most software. An advantage of multiple imputations over single imputation and complete case methods is that multiple imputations are flexible and can be used in a wide variety of scenarios. Multiple imputations can be used in cases where the data is missing completely at random and even when the data is missing not at random. However, the primary method of multiple imputations is multiple imputations by chained equations (MICE). It is also known as "fully conditional specification" and, "sequential regression multiple imputations" (Wulff and Ejlskov, 2017).

MICE is a popular adaption of missing imputation and is available to the user through the most commonly used software packages. MICE changes the imputation problem to a series of estimations where each variable takes its turn in being regressed on the other variables (Kaiser, 2014). MICE loops through the variables predicting each variable dependent on the others. This procedure provides excellent flexibility as each variable can be assigned a suitable distribution, e.g., poisson, linear or binomial (Wulff and Ejlskov, 2017).

Another notable local approach is the MICE-CART, which consists of multiple imputations by chained equations (MICE) and classification and regression trees (CART). It is a nonparametric approach made to perform multiple imputations through chained equations using sequential regression trees as the conditional models (Moorthy et al., 2014). In CART methodology, the best split is found over all possible splits generated by all predictors, which minimises the impurity of the response variable within the two sub-nodes where the impurity is a measure of deviance or variation for a numerical response (in regression trees) and a measure of heterogeneity or entropy for a categorical response (in classification trees) (Edwards and Finch, 2018).

4.2 Clustering

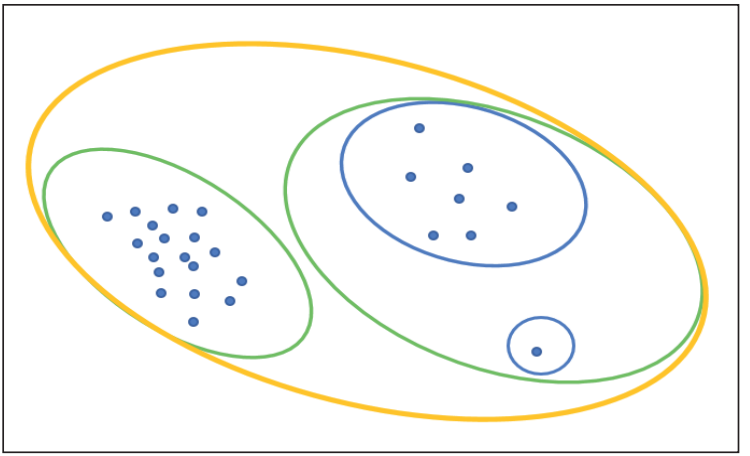
Clustering is an essential data mining tool for the analysis of Big Data and aims to consolidate the significant class data objects (clusters) so that objects grouped in the same cluster are similar and consistent according to specific parameters (Zerhani et al., 2015). The task is to arrange a set of objects so that the objects in the identical group are more related to each other than to those in other groups (clusters). Clustering belongs to unsupervised learning. Clustering algorithms can be classified into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms (Chitra and Maheswar, 2017).

Hierarchical clustering

Hierarchical clustering is a recursive partitioning of a dataset into successively smaller clusters. The input is a weighted graph where the edge weights represent pairwise similarities or dissimilarities between data points (Tan et al., 2018). Hierarchical

clustering is represented by a rooted tree where each leaf represents a data point and each internal node represents a cluster containing its descendant leaves. Computing a hierarchical clustering is a fundamental problem in data analysis; it is routinely used to analyse, classify, and pre-process large datasets (Cohen-Addad et al., 2018). There is extensive literature available on hierarchical clustering and its applications although it is impossible to discuss most of it in this paper. For some applications, the reader may refer to, e.g., (Hubert, 1977; Felsenstein, 2003; Castro et al., 2004).

Figure 1 | Hierarchical clustering



Source: Authors' own processing

The key operation of this algorithm is the computation of the proximity between two clusters, and it is the definition of cluster proximity that differentiates the various agglomerative hierarchical techniques that we will discuss. Cluster proximity is typically defined with a particular type of cluster in mind. Many agglomerative hierarchical clustering techniques come from a graph-based view of clusters (Rani and Rohil, 2013).

Determining the number of clusters

For determining the number of clusters, we will use the McClust method where the number of mixing components and the covariance parameterisation are selected using the Bayesian Information Criterion (BIC). In one dimension, there are just two models: E for equal variance and V for varying variance. In the multivariate setting, the volume, shape, and orientation of the covariances can be constrained to be equal or variable across groups. Thus, fourteen possible models can be specified (Scrucca et al., 2016).

Figure 2 | BIC models

Model	$\sum_k$	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	Equal	Equal	–
VII	$\lambda_k I$	Spherical	Variable	Equal	–
EEI	λA	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k A$	Diagonal	Variable	Equal	Coordinate axes
EVI	λA_k	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	λDAD^T	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda DA_k D^T$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k DAD^T$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k DA_k D^T$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda D_k AD^T{}_k$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k D_k AD^T{}_k$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda D_k A_k D^T{}_k$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k D_k A_k D^T{}_k$	Ellipsoidal	Variable	Variable	Variable

Source: Scrucca et al. (2016; edited)

Ward's method

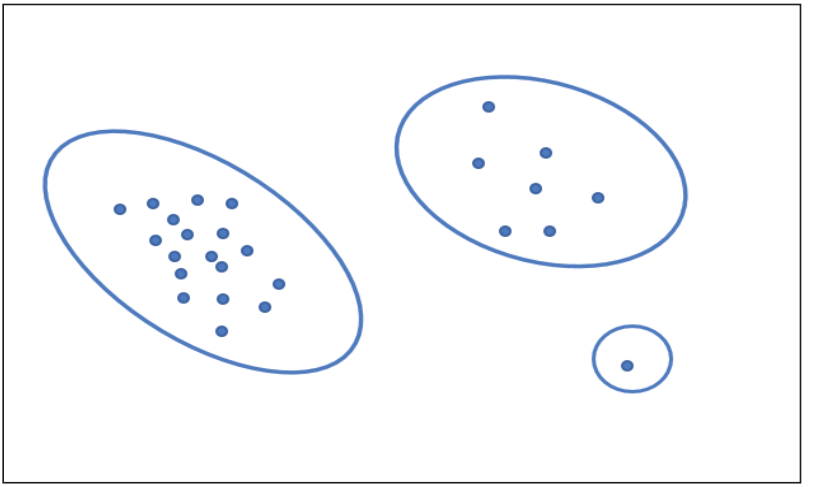
We can also take a prototype-based view, in which each cluster is represented by a centroid. The centroid method uses the centroid (centre of the group of cases) to determine the average distance between clusters of cases. An alternative technique to the usual centroid method is Ward's method. This method assumes that a cluster is represented by its centroid, but it measures the proximity between two clusters in terms of the increase in the SSE (squared error) that results from merging the two clusters. Similar to K-means, Ward's method attempts to minimise the sum of the squared distances of points from their cluster centroids (Tan et al., 2018).

Partitional clustering

Partitional clustering is the most popular class of clustering algorithm and is also known as an iterative relocation algorithm. These algorithms minimise a given clustering criterion by iteratively relocating data points between clusters until an optimal partition is attained (Chitra and Maheswar, 2017). A partitioning clustering algorithm splits the data points into k division, where each division represents a cluster and , where n is the number of data points. Partitioning methods are based on the idea that a cluster can be represented by a centre point. The partition is based on a certain objective function. The clusters are formed to optimise an objective partitioning criterion, such as a dissimilarity function

based on distance, so that the objects within a cluster are "similar", whereas the objects in different clusters are "dissimilar". Partitioning clustering methods are useful for applications where a fixed number of clusters are required. K-means, PAM (Partition around mediods) and CLARA are some of the partitioning clustering algorithms (Popat et al., 2014).

Figure 3 | Partitional clustering



Source: Authors' own processing

K-means clustering

K-means is one of the most popular partition-based methods and partitions the dataset into *k* disjoint subsets, where *k* is predetermined. The algorithm keeps adjusting the assignment of the objects to the closest current cluster mean until no new assignments of objects to clusters can be made (Elavarasi et al., 2011). One advantage of this algorithm is its simplicity. It also has several drawbacks. It is very difficult to specify the number of clusters in advance. Since it works with squared distances, it is also sensitive to outliers. Another drawback is that the centroids are not meaningful in most problems (Popat et al., 2014). In this algorithm, a cluster is represented by its centroid, which is a mean (average) of the points within a cluster. This only works efficiently with numerical attributes and can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. The technique aims to partition *n* observations into *k* clusters in which every observation belongs to the cluster with the nearby mean (Chitra and Maheswar, 2017).

The K-means algorithm has several significant properties, such as high effectivity in dealing with huge data sets, and it only works with numeric values; the resulting clusters have convex shapes and this method frequently terminates at a local optimum, and not the global optimum, which is also one of the major disadvantages of this method. Another fact that can be considered as a disadvantage, namely that this algorithm can be used only when the mean of the data set is defined and requires specifying *k*, the number of clusters, in advance (Vijayalakshmi and Devi, 2012).

For the K-means method, there are several specific types of functions for measuring the distance between clusters. The most common is a Euclidean distance, which computes the root of the square differences between the coordinates of a pair of objects, as follows:

$$D_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}. \quad (2)$$

The Manhattan distance, or city block distance, represents the distance between points in a city road grid. It computes the absolute differences between the coordinates of a pair of objects (Grabusts, 2011). There are also other methods for distance measuring, such as the Minkowski, Cosine and Chebyshev functions (Bora and Gupta, 2014).

There are many articles concerning the K-means method application topic, such as *Constrained K-means Clustering with Background Knowledge* (Wagstaff et al., 2001), *Improving the Accuracy and Efficiency of the K-means Clustering Algorithm* (Nazeer and Sebastian, 2009) or *An Algorithm for Online K-Means Clustering* (Liberty et al., 2016) and many others. There is also an interesting option of Merging K-means with hierarchical clustering for identifying general shaped groups proposed by Peterson et al. (2018) although this is not our aim at this time.

5. Data Preparation

Almost 5% of the missing data points can be found in the dataset because some sensors did not work properly. First, it is necessary to choose a method and make an imputation of missing values to the dataset. During data imputation processing, multicollinearity in the dataset was found. For localisation and deleting the collinear variable, we used the VIF method. Fifty-nine variables were perfectly correlated, so they had to be deleted because they give the same information as the other variables present in the data file. In the table below (Table 1) are the basic statistics of the counted VIF for each variable. There is the lowest and highest value of VIF, median, average and quartiles.

Table 1 | Descriptive statistics of VIF

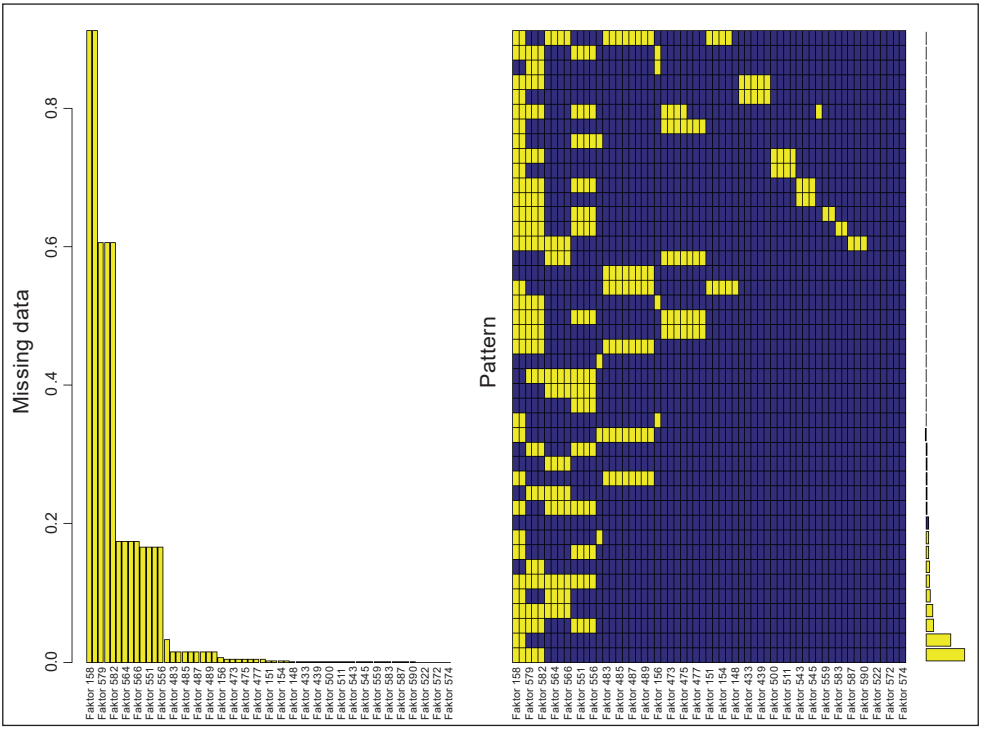
Minimum	Maximum	Median	Average	Q_1	Q_2	Q_3
1.20	2106285000	3.25	1254	1.66	3.25	10.61

Source: Authors` own processing

All variables where VIF is greater than five can be explained by other variables, which means that they can be deleted. After this data cleaning process, we obtained the remaining 55 variables, which can be reasonably included in the model.

After deletion of collinear and constant variables, we can proceed to the missing values imputation. In our dataset is the random missing data, as you can be seen in Figure 4.

Figure 4 | Missing data pattern



Source: Authors' own processing (RStudio)

For the missing data imputation, we have chosen the MICE-CART function, which is more accurate than the simple imputation of the mean, median or constant value. MICE-CART improves upon the standard MICE approach by automatically accounting for interaction effects among the variables for which imputation is needed (Moorthy et al., 2014). Now, there is the full dataset without collinear and constant variables, so the clustering analysis can begin.

6. Clustering Analysis

First, the clustering method is applied to the full dataset to recognise those products which passed quality control and the ones that failed. In this case, we want to have two clusters because we need to separate the products that passed from those that failed. The number of clusters intended is predetermined, so we will use the K-means method. After we determine the group of failed products, the hierarchical clustering method will be used for further analysis of the location of the origin of the defects.

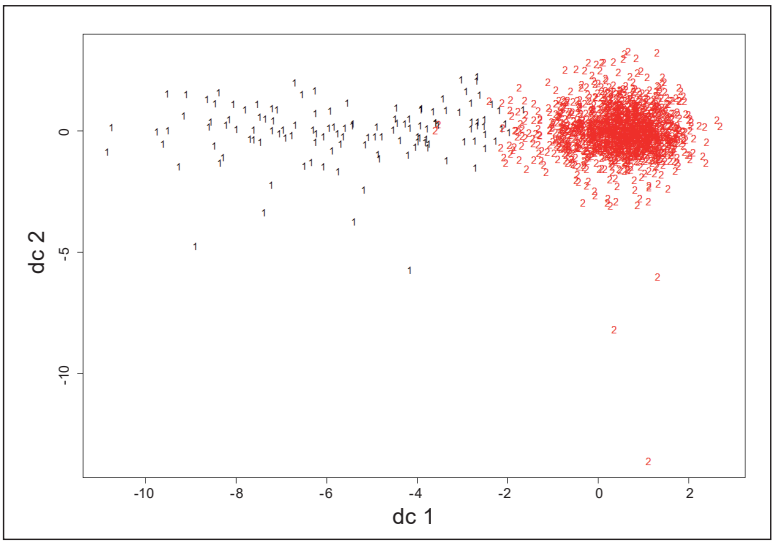
6.1 K-Means Clustering for All Products

Applying the K-means clustering method produces two significant clusters, as can be seen in Figure 5. The two components on the axes in the plot are the result of applying

the principal component analysis to the data. These are linear combinations of the input variables, which account for most of the variability of the observations.

We assume that the smaller (black) one represents the group of failed products evaluated on the quality control station. The second (red) one represents the group of products that are correct. According to this model, there are 1486 flawless products and 81 defective products. For the subsequent analysis, only the set of defective products will be used to identify groups of defect products with similar properties.

Figure 5 | K-means clustering for all products



Source: Authors` own processing (RStudio)

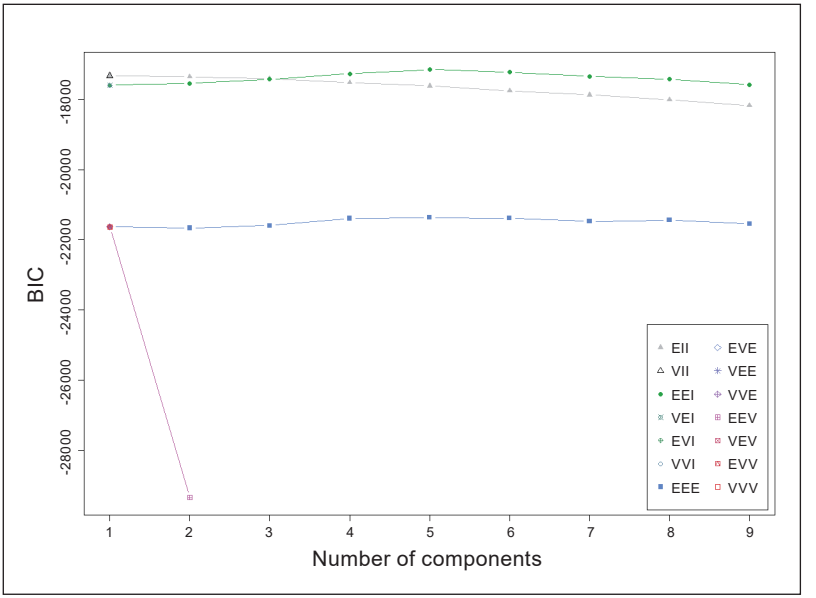
6.2 Hierarchical Clustering for Failed Products

The set of failed products was determined from the previous analysis; now, we will identify the groups of products with similar parameters by analysing only the failed products in order to recognise which products have similar defects. Through consecutive analysis directly in the production process, the results can be used for easier detection of the point in the process where the defects or the potential causes of the defects occur.

This time, we will use the hierarchical clustering analysis because we first need to determine the number of clusters that will be created.

In the following graph, we determine the optimal model and number of clusters according to the Bayesian Information Criterion for expectation-maximisation, initialised by hierarchical clustering for parameterised Gaussian mixture models. The plot showing the BIC traces (see Figure 6) for all the models is considered. We adjusted the range of the y-axis to remove those models with lower BIC values. There is a clear indication of the best option that is rendered by the EEI curve, according to the shape of which, we determine that the optimal number of clusters is five.

Figure 6 | Number of clusters determining



Source: Authors' own processing (RStudio)

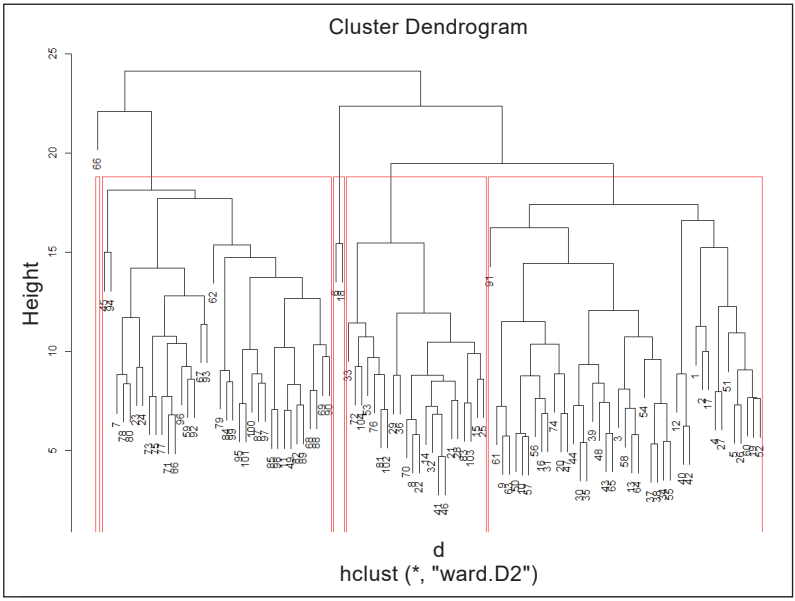
The hierarchical clustering method is now applied to five clusters using the Ward method and Euclidean distance measuring. The following dendrogram (Figure 7) shows the solution of this analysis where five clusters have been created. The products were grouped according to their parameters, the values which have been measured in the different stage of the production process. The smallest two clusters can only be inaccuracies in measuring or due to random employee mistakes. These defects will be difficult to analyse and will result in small costs for the company, so it is unnecessary to search now for their cause.

The other 3 clusters are of more interest to us. The defects to the products in these clusters are probably caused by the same event in the production process. Such an event may be, for example, bad settings on the machine, human failure or defects in the material used. These errors in the production process can cause huge additional costs for the organisation or loss of profit or market position. To find the exact cause of these defects, it is necessary to analyse the production process and map the material and resources flow.

Three considerable clusters of defective products with similar features or parameters appeared. At this point, it would be necessary to conduct an analysis of the production process but as the dataset was created by someone else, it makes it impossible. For this reason, we can only estimate the cause of the defects that arise.

From the resulting graph (Figure 7) it appears that there could be a connection between the products in the clusters due to their serial numbers. It is possible that a specific event in the production process can occur, such as machine failure, which could cause a few consecutive defect products. To determine particular causes, we would need more information about the production process, a record of machine failures, material quality review etc.

Figure 7 | Hierarchical clustering for failed products



Source: Authors` own processing (RStudio)

7. Conclusion

In the presented paper, we analysed the data concerning the scrap in semiconductor manufacturing in order to identify the main types and causes of defects in manufactured products. For this task, we used VIF and MICE-CART methods for data pre-processing (deletion of factors causing multicollinearity, imputation) and cluster analysis (K-means, hierarchical clustering).

Using the above-mentioned approaches, we detected 81 defective products from the total of 1567 products examined. The defective products have been divided into five clusters according to their similar properties. From the results of the hierarchical clustering analysis, it is obvious that there are three substantial sources of defects in the production process. We assume that the products in these groups have the same or similar cause of error. For closer investigation of the cause, it will be necessary to analyse the production process itself. We would need more information about the production process such as mapping of material and resources flows, records of machine failures or a material quality review. The other two clusters are insignificant because they are too small. These defects can be caused by a random event or human resource failure; the search for their cause would probably be more demanding than the potential cost savings made when implementing the corrective action. We have also proved that a combination of two different clustering algorithms in a sequence is possibly an effective and successful method of identifying and classifying the defects in the manufacturing process.

The limitations in this research are specifically the nature and the source of the data. In this case, the dataset came from the public source, so the supplementary information

about the production environment was very limited. Thus, the quality of the dataset can also be considered as a limitation. We had to perform data imputation and despite the nontrivial method of data, the results of the analysis could have been influenced by this. For the further improvement of accuracy of the research, it is theoretically possible to contact the authors of the dataset and obtain more details about the dataset, the manufacturing company involved and its processes.

References

AS/NZS ISO (1994). ISO 9001:1994 Quality Systems – Model for Quality Assurance in Design, Development, Production, Installation and Servicing, [online]. Available at: https://www.saiglobal.com/pdftemp/previews/osh/as/as9000/9000/9001.pdf [Accessed 25 Jun. 2018]

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In J. Kogan, C. Nicholas and M. Teboulle, eds., *Grouping Multidimensional Data* (pp. 25–71). Berlin: Springer.

Bhuvaneswari, S., and Sabarathinam, J. (2013). Defect Analysis Using Artificial Neural Network. *International Journal of Intelligent Systems and Applications*, 5(5), pp. 33–38. https://doi.org/10.5815/ijisa.2013.05.05

Bora, D. J., and Gupta, A. K. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab. *International Journal of Computer Science and Information Technologies*, 5(2), pp. 2501–2506.

Castro, R. M., Coates, M. J., and Nowak, R. D. (2004). Likelihood Based Hierarchical Clustering. *IEEE Transactions on Signal Processing*, 52(8), pp. 2308–2321. https://doi.org/10.1109/TSP.2004.831124

Cohen-Addad, V. et al. (2018). Hierarchical Clustering: Objective Functions and Algorithms. SODA (Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms). https://doi.org/10.1137/1.9781611975031.26

Crosby, P. B. (1979). *Quality Is Free: The Art of Making Quality Certain.* New York: McGraw-Hill.

Deming, W. E. (1982). *Quality Productivity and Competitive Position*. Cambridge, MA: MIT Press.

Dragulanescu, I.-V., and Popescu, D. (2015). Quality and Competitiveness: A Lean Six Sigma Approach. *Amfiteatru Economic Journal*, 17(9), pp. 1167–1182.

Edwards, J. M., and Finch, W. H. (2018). Recursive Partitioning Methods for Data Imputation in the Context of Item Response Theory: A Monte Carlo Simulation, *Psicológica*, 39, pp. 88–117. https://doi.org/10.2478/psicolj-2018-0005

Elavarasi, S. A., Akilandeswari, J., and Sathiyabhama, B. (2011). A Survey on Partition Clustering Algorithms. *International Journal of Enterprise Computing and Business Systems*, 1(1).

Evans, J. R. (2015). Modern Analytics and the Future of Quality and Performance Excellence. *Quality Management Journal*, 22(4), pp. 6–17. https://doi.org/10.1080/10686967.2015.11918447

Felsenstein, J. (2003). *Inferring Phylogenies* (2nd ed.). Oxford: Sinauer Associates.

Gaub, H. (2016). Customization of Mass-Produced Parts by Combining Injection Molding and Additive Manufacturing with Industry 4.0 Technologies. *Reinforced Plastics*, 60(6), pp. 401–404. https://doi.org/10.1016/j.repl.2015.09.004

Grabusts, P. (2011). Distance Metrics Selection Validity in Cluster Analysis. *Scientific Journal of Riga Technical University*, 45(1), pp. 72–77. https://doi.org/10.2478/v10143-011-0045-y

Horvath, M., and Vircikova, E. (2012). Data Mining Model for Quality Control of Primary Aluminum Production Process. *Management and Production Engineering Review*, 3(4), p. 47. https://doi.org/10.2478/v10270-012-0033-x

Hoyle, D. (1994). *ISO9000 – Quality Systems Handbook* (2nd ed.) Birlingham: Butford Technical Publishing.

Hubert, L. (1977). A Set-Theoretical Approach to the Problem of Hierarchical Clustering. *Journal of Mathematical Psychology*, 15(1), pp. 70–88. https://doi.org/10.1016/0022-2496(77)90041-4

Chammas, A. et al. (2015). Drift Detection and Characterization for Condition Monitoring: Application to Dynamical Systems with Unknown Failure Modes. *IMA Journal of Management Mathematics*, 26, pp. 225–243. https://doi.org/10.1093/imaman/dpu008.

Chitra, A., and Maheswari, D. (2017). A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Computer Science and Mobile Computing*, 6(8), pp. 109–115.

Choudhary, A. K., Tiwari, M. K., and Harding, J. A. (2009). Data Mining in Manufacturing: A Review Based on the Kind of Knowledge. *Journal of Intelligent Manufacturing*, 20(5), pp. 501–521. https://doi.org/10.1007/s10845-008-0145-x

Kaiser, J. (2014). Dealing with Missing Values in Data. *Journal of Systems Integration*, 5(1), pp. 42–51. https://doi.org/10.20470/jsi.v5i1.178

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, NJ: Wiley-Interscience.

Kao, H.-A. et al. (2017). Quality Prediction Modeling for Multistage Manufacturing Based on Classification and Association Rule Mining. *MATEC Web of Conferences*, 123. https://doi.org/10.1051/matecconf/201712300029

Kerdprasop, K., and Kerdprasop, N. (2011). Feature Selection and Boosting Techniques to Improve Fault Detection Accuracy in the Semiconductor Manufacturing Process, [online]. *IMECS*. Available at: http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp398-403.pdf [Accessed 25 Jul. 2018]

Knowles, G. (2011). *Quality Management*, [online]. Available at: http://www.znrfak.ni.ac.rs/SERBIAN/010-STUDIJE/OAS-3-2/PREDMETI/III%20GODINA/316-KOMUNALNI%20SISTEMI%20I%20ZIVOTNA%20SREDINA/SEMINARSKI%20RADOVI/2014/S175%20-%20S200.pdf [Accessed 25 Jun. 2018]

Köksal, G., Batmaz, I., and Testik, M. C. (2011). A Review of Data Mining Applications for Quality Improvement in Manufacturing Industry. *Expert Systems with Applications*, 38. https://doi.org/10.1016/j.eswa.2011.04.063

Kutner, M. H. et al. (2005). *Applied Linear Statistical Models* (5th ed.). New York: McGraw-Hill.

Liberty, E., Sriharsha, R, and Sviridenko, M. (2015). An Algorithm for Online K-Means Clustering. *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pp. 81–89. https://doi.org/10.1137/1.9781611974317.7

Mabkhot, M. M. et al. (2018). Requirements of the Smart Factory System: A Survey and Perspective. *Machines*, 6(2), p. 23. https://doi.org/10.3390/machines6020023.

McCann, M., and Johnston, A. (2008). *SECOM Data Set*, *Index of /ml/machine-learning-databases/secom*. Available at: https://archive.ics.uci.edu/ml/machine-learning-databases/secom/secom.names [Accessed 13 Apr. 2018]

Mehran, E., and Mehran, S. (2013). Quality Management and Performance: An Annotated Review. *International Journal of Production Research*, 51(18), pp. 5625–5643. https://doi.org/10.1080/00207543.2013.793426

Moorthy, K., Mohamad, M. S., and Deris, S. (2014). A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data. *Current Bioinformatics*, 9(1), pp. 18–22. https://doi.org/ 10.2174/15748936089991401091209 57

Nazeer, K. A. A., and Sebastian, M. P. (2009). Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm. [online]. *Proceedings of the World Congress on Engineering*, 1. Available at: www.iaeng.org/publication/WCE2009/WCE2009_pp308-312.pdf [Accessed 12 Apr. 2018]

Paul, R. K. (2006). Multicollinearity: Causes, Effects and Remedies, [online]. Available at: https://www.researchgate.net/publication/255640558_MULTICOLLINEARITY_CAUSES_EFFECTS_AND_REMEDIES [Accessed 25 Apr. 2018]

Peterson, A. D., Ghosh, A. P., and Maitra, R. (2018). Merging K-Means with Hierarchical Clustering for Identifying General-Shaped Groups. *Stat (The ISI's Journal for the Rapid Dissemination of Statistics Research)*, 7(1). https://doi.org/10.1002/sta4.172

Popat, S. K. et al. (2014). Review and Comparative Study of Clustering Techniques. *International Journal of Computer Science and Information Technologies*, 5(1), pp. 805–812.

Ramana, E. V., and Reddy, P. R. (2013). Data Mining Based Knowledge Discovery for Quality Prediction and Control of Extrusion Blow Molding Process. *International Journal of Advances in Engineering & Technology*, 6(2), pp. 703–713.

Rani, Y., and Rohil, H. (2013). A Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology*, 3(11), pp. 1225–1232.

Rokach, L., Romano, R., and Maimon, O. (2008). Mining Manufacturing Databases to Discover the Effect of Operation Sequence on the Product Quality. *Journal of Intelligent Manufacturing*, 19(3), pp. 313–325. https://doi.org/10.1007/s10845-008-0084-6

Saad, N. H. et al. (2015). Defect Segmentation of Semiconductor Wafer Image Using K-Means Clustering. *Applied Mechanics and Materials*, 815, pp. 374–379. https://doi.org/10.4028/www.scientific.net/AMM.815.374

Sabet, S. A. A. M., Moniri, A., and Mohebbi, F. (2017). Root-Cause and Defect Analysis Based on a Fuzzy Data Mining Algorithm. *International Journal of Advanced Computer Science and Applications*, 8(9), pp. 21–28. https://doi.org/10.14569/IJACSA.2017.080903

Sadikoglu, E., and Zehir, C. (2010). Investigating the Effects of Innovation and Employee Performance on the Relationship between Total Quality Management Practices and Firm Performance: An Empirical Study of Turkish Firms. *International Journal of Production Economics*, 127(1), pp. 13–26. https://doi.org/10.1016/j.ijpe.2010.02.013

Saludes-Rodil, S., Baeyens, E., and Rodríguez-Juan, C. P. (2015). Unsupervised Classification of Surface Defects in Wire Rod Production Obtained by Eddy Current Sensors. *Sensors*, 15(5), pp. 10100–10117. https://doi.org/10.3390/s150510100

Schumacher, A., Erol, S., and Sihn, W. (2016). A Maturity Model for Assessing Industry 4.0 Readiness and Maturity of Manufacturing Enterprises. *Procedia CIRP*, 52, pp. 161–166. https://doi.org/10.1016/j.procir.2016.07.040

Scrucca, L. et al. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1), pp. 289–317.

Tan, P.-N. et al. (2018). Cluster Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining*, 2nd ed. (pp. 525–611). London: Pearson.

Tseng, T.-L., Jothishankar, M. C., and Wu, T. (2004). Quality Control Problem in Printed Circuit Board Manufacturing – An Extended Rough Set Theory Approach. *Journal of Manufacturing Systems*, 23(1), pp. 56–72. https://doi.org/10.1016/S0278-6125(04)80007-4

Vijayalakshmi, M., and Devi, R. (2012). A Survey of Different Issue of Different Clustering Algorithms Used in Large Data Sets. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(3), pp. 305–307.

Wagstaff, K., and Cardie, C. (2001). Constrained K-Means Clustering with Background Knowledge, [online]. *Proceedings of the Eighteenth International Conference on Machine Learning*. Available at: https://pdfs.semanticscholar.org/0bac/ca0993a3f51649a6bb8dbb093fc8d8481ad4.pdf [Accessed 14 Apr. 2018]

Wang, K. (2006). Data Mining in Manufacturing: The Nature and Implications. In K. Wang, G. L. Kovacs, M. Wozny and M. Fang, eds., *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management* (pp. 1–10). Proceedings of PROLAMAT 2006, IFIP TC5 Conference, June 15–17, 2006, Shanghai, China. Boston, MA: Springer. https://doi.org/10.1007/0-387-34403-9_1

Wang, K.-S. (2013). Towards Zero-Defect Manufacturing (ZDM) – A Data Mining Approach. *Advances in Manufacturing*, 1(1), pp. 62–74. https://doi.org/10.1007/s40436-013-0010-9

Wulff, J., and Ejlskov, L. (2017). Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. *The Electronic Journal of Business Research Methods*, 15(1), pp. 41–56.

Yiakopoulos, C. T., Gryllias, K. C. and Antoniadis, I. A. (2011). Rolling Element Bearing Fault Detection in Industrial Environments Based on a K-Means Clustering Approach. *Expert Systems with Applications*, 38(3), pp. 2888–2911. https://doi.org/10.1016/j.eswa.2010.08.083

Yin, T. S. et al. (2018). Comparing Quality Management Practices between Food Industry and Electrical and Electronic Industry. In V. Ribiere, ed., *Proceedings of the International Conference on Management, Leadership & Governance*, Bangkok, 24–25 May (pp. 326–332). Reading, UK: Acad. Conf. and Publish. International Limited.

Yusof, M. et al. (2017). Clustering of Frequency Spectrums from Different Bearing Fault Using Principle Component Analysis. *MATEC Web of Conferences*, 90. https://doi.org/10.1051/matecconf/20179001006

Zerhari, B., Lahcen, A. A., and Mouline, S. (2015). Big Data Clustering: Algorithms and Challenges [conference paper, online]. *International Conference on Big Data, Cloud and Applications BDCA'15*, Tetuan, Morocco. Available at: https://www.researchgate.net/publication/276934256_Big_Data_Clustering_Algorithms_and_Challenges [Accessed 15 Apr. 2018]