Mould wear-out prediction in the plastic injection moulding industry
a case study

Frumosu, Flavia Dalia; Rønsch, Georg Ørnskov; Kulahci, Murat

Published in:
International Journal of Computer Integrated Manufacturing

Link to article, DOI:
10.1080/0951192X.2020.1829062

Publication date:
2021

Document Version
Peer reviewed version

Mould wear-out prediction in the plastic injection moulding industry: A case study

Flavia Dalia Frumosu[a], Georg Ørnskov Rønsch[a], and Murat Kulahci[a,b]

[a]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark; [b]Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

ABSTRACT

The current work addresses an industrial problem related to injection moulding manufacturing with focus on mould wear-out prediction. Real data sets are provided by an industrial partner that uses a multitude of moulds with different shapes and sizes in its production. An analysis of the data is presented and begins with clustering the moulds based on their characteristics and pre-chosen running settings. Using the results of the clustering, the mould wear-out is modelled using Kaplan-Meier survival curves. Furthermore, a random survival forest model is fitted for comparison and model performance is assessed. The main novelty of the case study is the implementation of mould wear-out prediction in real-time with the outcomes presented in terms of conditional survival curves including a proposed early warning system. For visualization and further industrial implementation, a R Shiny dashboard is developed and presented.

KEYWORDS

Industry 4.0; predictive maintenance; injection moulding; mixed data; reliability analysis; censored data; mould wear-out

Introduction

Like many industries, the plastic injection moulding industry is currently going through changes under the so-called Industry 4.0 (Lasi et al. 2014), which aims for higher degree of automation and digitalization. Companies are already embarked on the route to Industry 4.0 and even national policies were elaborated for a number of countries (Da Silva et al. 2020). With increasing availability of sensors and decreasing cost of data storage and computer power, this transformation often results in abundance of process data for both offline and online data analytics. For the latter, one specific area of application is in Condition-Based-Maintenance (CBM) (Peng, Dong, and Zuo 2010). CBM is a maintenance program in which maintenance decisions are based on information from condition monitoring and it consists of data acquisition, data processing and maintenance decision-making (Jardine, Lin, and Banjevic 2006). This is a significant departure from corrective maintenance (CM) in which the maintenance is performed after the failure happens, and from preventive maintenance where the maintenance is performed on regularly scheduled intervals irrespective of the condition of the production equipment (Ahmad and Kamaruddin 2012). Significant attention has been paid to CBM in the literature and more recently, the focus has been shifted towards predictive maintenance (PdM) (Li, Wang, and Wang 2017). According to Li, Wang, and He (2016), the goal of PdM is to reduce downtime and cost of maintenance through the monitoring of the equipment's working conditions as well as predicting equipment failure that allows the maintenance to be planned before the actual fault occurs. There is a strong connection between CBM and PdM, where CBM can be considered as PdM (Sai, Shcherbakov, and Tran 2019; Hashemian 2010) or CBM can be treated as an effective form of PdM (Amari, McLaughlin, and Pham 2006).

Recent work related to PdM within Industry 4.0 includes Li, Wang, and Wang (2017) who propose a system

CONTACT Flavia Dalia Frumosu, fdal@dtu.dk, Technical University of Denmark, Department of Applied Mathematics and Computer Science, Richard Petersens Plads, Building 324, DK - 2800 Kgs. Lyngby Denmark.

framework built on Industry 4.0 concepts that includes a fault analysis process and treatment used for predictive maintenance in machine centers. Cachada et al. (2018) propose an intelligent and predictive maintenance system along with its architecture that is aligned with Industry 4.0 principles. Haarman, Mulders, and Vassiliadis (2017) have introduced the concept of Predictive Maintenance 4.0 (PdM 4.0) which is about the prediction of future failures in assets and the selection of the most effective preventive measure through the means of advanced analytic techniques applied on big data.

The case study addressed in this article is from the plastic injection moulding industry with the ultimate aim of mould wear-out prediction. In this regard, it is related to a study of Remaining Useful Life (RUL) of an asset or system. RUL can be defined as the time between the current time until the end of the useful life, Si et al. (2011), where the emphasis is set on statistical methods for RUL estimation based on both directly and indirectly observed state processes. RUL estimation is one of the key issues in CBM as shown in Jardine, Lin, and Banjevic (2006), Peng, Dong, and Zuo (2010) as well as in Ahmad and Kamaruddin (2012). RUL estimation is an important pillar for predictive maintenance platforms (Aivaliotis, Georgoulias, and Chryssolouris 2019).

With the increasing availability of production data that also includes maintenance data, there is currently more focus on data driven approaches to maintenance planning and scheduling. Bukkapatnam et al. (2019) proposes Manufacturing System-wide Balanced Random Survival Forest (MBRSF), a nonparametric machine learning approach for long-term prognosis for breakdowns of production equipment. Alsina et al. (2018) compares machine learning methods for predicting component reliability. Ragab et al. (2016) presents a prognostic methodology based on Kaplan – Meier estimation for RUL.

The case study is based on several data sets provided by an industrial partner. The data that is subject to changes in time is denoted as time dependent and data that remains constant in time is denoted as time independent. Mould characteristics and pre-chosen running settings are examples of time independent data whereas production and maintenance data are time dependent. In this work, the mould pre-chosen running settings are defined as machine settings selected by the operator prior to the use of the mould in production. Furthermore, it is assumed that these settings are not changed during production.

Mould wear-out prediction is the main goal as it has a direct impact on product quality. Moreover, understanding the deterioration mechanisms of moulds is expected to facilitate effective maintenance planning and hence, reduce the cost of maintenance. The available data contains information on a limited number of worn-out moulds. The proposed solution and choice of model reflect this specific characteristic of the data. However, the authors also provide a data intensive modeling approach that can be implemented as data for more worn-out moulds becomes available.

The main research contributions of the article are in terms of prediction and monitoring of mould wear-out. A method for prediction of mould wear-out is developed based on conclusions drawn from the analysis of real data. Moreover, a monitoring strategy in the shape of an early warning system, which can be used by the industrial partner by means of a dashboard, is presented.

The first part of the study focuses on dimensionality reduction of the time independent data so that it is easier for practitioners to interpret and monitor the results. The second part of the study focuses on the prediction of mould wear and tear. Product quality data and more sensor data will only be available in the future. However, a prediction model using this additional data is also discussed. Throughout the study a special emphasis is put on visualization and interpretation. As the final product is meant for industrial use, the results are presented in the shape of a R Shiny (2019) dashboard. Finally, future directions for research and application are discussed. It should be noted that due to confidentiality reasons, the data and the results have been masked when considered necessary.

1. Industrial context

In this section, the technical background for the injection moulding process is provided as well as the description of various data sets used in this study.

Injection moulding process

The basic mechanism of an injection moulding machine is presented in Figure 1.

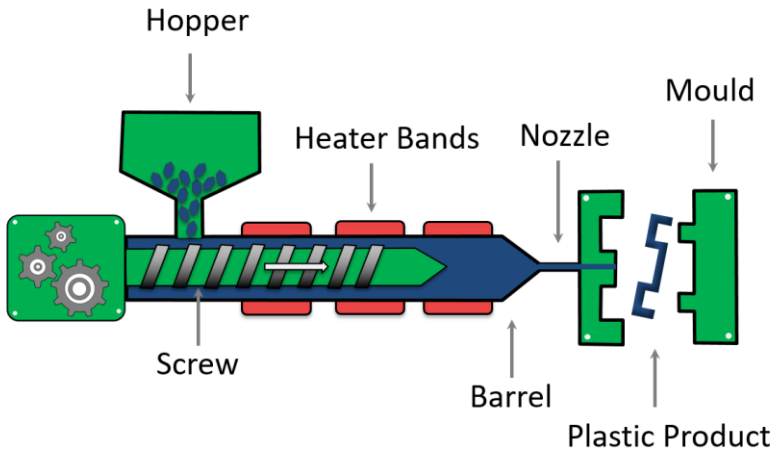

Figure 1. Injection moulding machine with major components.

The injection moulding production is discussed in terms of "shots" during which the moulding of a product (or a group of products depending on the mould characteristics) is performed. First, plastic pellets are fed into the hopper and the barrel. The plastic pellets in the barrel start to melt gradually due to the rotation of the screw that generates shear heat and due to external heat that is provided by the heater bands wrapped around the barrel. Next, as the screw rotates, the molten plastic is pushed into the nozzle which sends further the molten plastic to the mould. Usually, moulds have more than one cavity which means that more products are manufactured per shot. The plastic products from the mould are cooled while the mould remains closed via cooling channels placed inside the mould. After the plastic products solidify, the mould is opened and the plastic products are ejected from the mould, Figure 1.

The injection moulding process is a complex process during which various pressures, temperatures and speed-readings are collected and used for controlling the process through engineering control. Furthermore, the optimal settings of these variables are set by the operator based on prior experience. Hence, the process is controlled through both automated and manual interventions.

Data description

The industrial partner provided different types of real data sets related to the moulds used in production. The moulds are used for the manufacturing of plastic parts of different shapes and sizes and are usually with more than one cavity. The worn-out condition of the mould is defined internally by the industrial partner considering, for example, visual inspections. However, for confidentiality reasons, the exact definition is not presented in the article. The data sets are connected through the mould ID which is unique for each mould and are described in Table 1. All the data sets with the current status "Available" are analyzed or utilized in the article.

The mould characteristics include, for example, number of cavities and mould dimensions while the mould pre-chosen running settings are machine settings selected by the operator before the mould is used in

production. Mould characteristics and pre-chosen running settings are not time dependent as these are assumed to be constant in time.

The maintenance data contains information regarding, for example, the number of cleanings, change of spare parts or other relevant maintenance information. Production data is related, for example, to the number of produced parts and stops in production. The maintenance and production data sets are time dependent as changes can occur over time.

The table also contains information about data sets that will become available in the future and can potentially have impact in determining the condition of the mould. Data sets containing both categorical data and numerical data are labeled as mixed data. The variables of the data sets in Table 1 are denoted as features, which is the machine learning terminology for variables. Furthermore, missing observations are present in some data sets as in the mould characteristics and running settings.

Table 1. Description of data sets.

Data set	Time dependence	Number Features	Data Type	Current status
Mould status (worn-out or running)	Dependent	2	numeric	Available
Mould characteristics	Independent	Around 20	mixed	Available
Mould pre-chosen running settings	Independent	Around 180	mixed	Available
Maintenance	Dependent	Around 10	mixed	Available
Production	Dependent	Around 15	numeric	Available
Environmental conditions	Dependent	N/A	N/A	Missing
Injection moulding machine sensors	Dependent	N/A	N/A	Missing
Metrology for products	Dependent	N/A	N/A	Missing

When determining the time for reaching a worn-out status, the number of shots is used as the unit of measurement. The number of shots provides a more realistic picture about the life of a mould compared to calendar time, as there are periods during which the mould is idle due to, for example, maintenance work. Following the terminology from survival analysis (Wang, Li, and Reddy 2019), the moulds that are currently still in use generate so-called right-censored data whereas the worn-out moulds generate uncensored. A depiction of the data is presented in Figure 2.

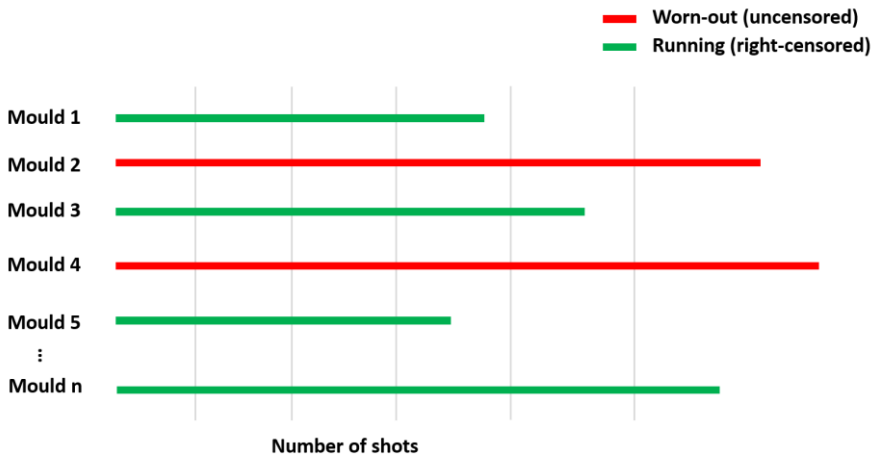

Figure 2. A depiction of worn-out versus running moulds.

Industrial problem

The primary goal of the case study is to assess the condition of the mould and predict when it will be worn-out. Another goal is to provide visual tools for the real time surveillance of the mould condition in production for the operators. It is important to note that the mould data includes a variety of different

suppliers and production sites. Therefore, it is also crucial for the industrial partner to have an overview of the moulds performance at a global level, where consistency is key. Hence, the first step in the analysis is the empirical clustering of the moulds expecting the clusters to be formed by different characteristics and pre-chosen production settings of the moulds. This is followed by the modeling of the moulds' condition for each cluster. The methodology followed in the article is presented in the framework from Figure 3.

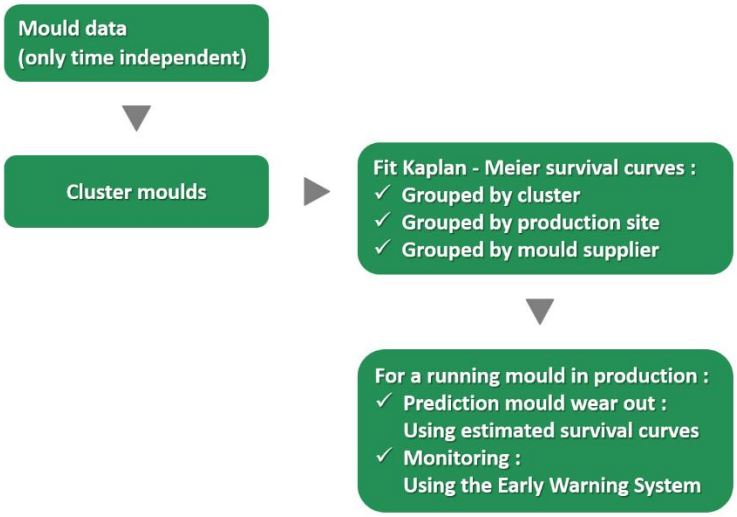

Figure 3. Framework describing the methodology followed in the article.

2. Theoretical background

Cluster analysis

Clustering is an unsupervised learning method which is used to group observations into clusters which are "similar" within the groups and "dissimilar" between groups (Ahmad and Khan 2019; Foss, Markatou, and Ray 2019). In our case, the data sets regarding the mould characteristics contain mixed data, which makes clustering challenging since applying direct mathematical operations such as summation or averaging used in obtaining similarity measures is not possible (Ahmad and Khan 2019). Mixed data is becoming more ubiquitous as modern production data often comes from increasingly heterogeneous sources (Foss, Markatou, and Ray 2019).

In general, clustering is performed using distance measures which quantify "(dis)similarity" between observations in the feature space (Foss, Markatou, and Ray 2019). For mixed data, a common distance measure is the so-called Gower distance (Gower 1971). After the distances between observations are computed through Gower or any other distance measure, clustering algorithms like k-means or hierarchical clustering (Hastie, Tibshirani, and Friedman 2013) can be applied.

In this study, an unsupervised random forest (Shi and Horvath 2006) is used for obtaining the distances between observations, which are called "proximities" (Breiman and Cutler 2019). The main advantages of the unsupervised random forest are its ability to deal with mixed data as well as allowing computation of variable importance, which can be of great value for interpreting the model outcome. In unsupervised random forest, for a given data set of N observations, the joint distribution of the variables is obtained and a synthetic data set is generated using this joint distribution. Often, the number of observations is kept the same for both the original and synthetic data sets. Each data set is labeled separately as for example 1 and 2, and this label is used as the response for the random forest model. Based on this model, an N x N similarity matrix is obtained through the frequency of any two observations, *i* and *j*, that are placed in the same terminal node. The variable importance can be computed using the Gini index decrease, i.e., the decrease in

node impurity (Hastie, Tibshirani, and Friedman 2013). Each time a node split is made on a feature *m*, the Gini index for two descendant nodes becomes less than that of the parent node (Breiman and Cutler 2019). By adding up the Gini index decreases for each individual feature over all trees, the variable importance can be computed for each feature.

For assessing the number of clusters, Partition Around Medoids (PAM) (Kaufman and Rousseeuw 1987) along with the average silhouette width (Rousseeuw 1987) is applied. The PAM algorithm is very similar to K-means since both are partitional algorithms. The main difference is that K-means works with centroids while PAM works with medoids, which have the property that the average dissimilarity to all of the observations in the cluster is minimized. Every cluster can be represented by a silhouette (Rousseeuw 1987) that is built on the comparison of the clusters' tightness and separation. The average silhouette width can then be used to select a suitable number of clusters, with a high average silhouette width indicating a good clustering. There are various methods to assess the right number of clusters, however, for the purpose of the case study this approach is considered sufficient.

For clustering, only the results of PAM could be used, however, a visual interpretation is of interest and for this it was decided to use t-distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton 2008). In t-SNE, the emphasis is put on modelling the dissimilar observations by large pairwise distances, while the similar observations are modelled by small pairwise distances. This approach was used as it provides a projection of the high-dimensional data into a two or three-dimensional map and thus, can be visualized in a plot. For the case study, a two-dimensional map was used.

Survival analysis

The results of the cluster analysis can be further used to check the worn-out and running moulds survival functions. As it was described in the Data description section and in Figure 2, the worn-out moulds can be regarded as uncensored data and the running moulds as right-censored data. Given this description, the life of the moulds can be presented in terms of a survival function (or curve) also called the reliability function. For the current article, the term survival curve is preferred.

If $T(\geq 0)$ is the number of shots until the mould reaches the worn-out state, then the survival function can be expressed as:

$$S(t) = P(T > t) \tag{1}$$

where, $S(t)$ is the probability that the mould is not worn-out by t number of shots.

Using the historical data, there are different ways of estimating the survival function in Equation (1). The Kaplan-Meier estimator (Kaplan and Meier 1958) is used:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2}$$

where t_i is a number of shots when at least one event happened, d_i is the total number of worn-out events at t number of shots and n_i is the total number of moulds at risk at t number of shots.

Using the Kaplan-Meier estimator, the survival curve for each of the clusters (categories) from the cluster analysis can be estimated. For testing if there is a statistical significant difference between two or more survival curves, the G^{ρ} family of tests with $\rho = 0$ is used which is the log-rank or Mantel-Haenszel test (Harrington and Fleming 1982).

For a running mould belonging to a certain cluster, the survival curve can be displayed as a conditional survival curve (Zabor et al. 2013; Hieke et al. 2015). This is more intuitive than displaying the entire survival curve from zero number of shots. The confidence intervals can also be displayed. For this, the same approach as in Zabor et al. (2013) has been used for the case study.

As discussed earlier, the available data sets also contain production and maintenance information

associated with the worn-out and running moulds. For building a model using this information as well, a random survival forest (Ishwaran et al. 2008) can be used for the case study. A random survival forest model is an ensemble learner based on the averaging of a tree base learner (Ishwaran et al. 2011). For the survival setting, the base-learner is formed from a binary survival tree and the ensemble is formed from a cumulative hazard function by averaging the Nelson-Aalen's cumulative hazard function for each tree (Ishwaran et al. 2011).

There are several advantages associated with a random survival forest model. As specified by Boström et al. (2018), a random survival forest model is robust for high-dimensional data and can accommodate high-level interactions between features. Other benefits inherited from random forests include the performance estimation using the Out-of-Bag (OOB) predictions, which can easily be implemented in parallel as well as variable importance. Furthermore, for survival analysis problems, random survival forest models do not require model assumptions such as proportional hazards, Ishwaran et al. (2008). In terms of the performance measure, the authors decided to use the c-index (also called as concordance or Harrell's index) (Harell et al. 1982). The c-index ranges from 0 to 1 and computes the proportion of concordant pair of observations among all pairs of observations in terms of survival time (Harrell, Lee, and Mark 1996).

Early warning system

Once the moulds are clustered, the next step in the case study is to provide an early warning system regarding the running mould status from cluster *c* at a given shot *t* ($S_t^{(c)}$) compared to historical data of the moulds from the same cluster. For this, the authors propose monitoring a percentage that is composed from the ratio of the worn-out moulds, which were still running at $S_t^{(c)}$ divided by the total number of worn-out moulds from that specific cluster. In this way, the operator can have a quick overview on how the running mould is behaving in relationship with previous moulds that were worn-out from the same cluster. If $S_i^{(c)}$ is defined as the number of shots at which the moulds in a certain cluster c were worn-out and taken off production, for $i = 1, \dots, n_c$, then that percentage can be expressed as:

$$p_{active} = \frac{\sum_{i=1}^{n_c} \delta_i}{n_c} \tag{3}$$

where, $\delta_i = \begin{cases} 1, & S_i^{(c)} \geq S_t^{(c)} \\ 0, & S_i^{(c)} < S_t^{(c)} \end{cases}$

Next steps involve the development of an early warning system using the maintenance and production data sets. The idea is that at $S_t^{(c)}$ the running mould should behave similarly in terms of maintenance and production with the worn-out moulds at $S_t^{(c)}$ that belong to the same category as the running mould.

For the online tracking, the authors propose a method for monitoring based on a Hotelling T^2 statistic (Montgomery 2012). For this purpose, at $S_t^{(c)}$, the average measurements is calculated for features of production and maintenance data of the worn-out moulds from first shot to $S_t^{(c)}$. This forms the basis to compare the running moulds at $S_t^{(c)}$. The data is denoted as W, which has the number of observations $m = \sum_{i=1}^{n_c} \delta_i$ and p features. The average is used since different features are collected at varying sampling frequencies. The same calculations are performed for a mould in use and the vector of size 1 x p is denoted as r. The Hotelling T^2 statistic is calculated as follows:

$$T^2 = (x - \bar{x})' K^{-1} (x - \bar{x}) \tag{4}$$

where, $\bar{x}$ is the sample mean vector and K is the covariance matrix obtained from W. This statistic gives the

"scaled distance" of the current mould to the cluster in which it belongs in terms of its production and maintenance data. A threshold is then set, T_{UL} beyond which the current mould may no longer be considered belonging to that particular cluster depicted in W. The threshold is calculated using:

$$T_{UL} = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,m-p} \tag{5}$$

Where, m is the number of samples, p is the number of variables/features and $F_{\alpha,p,m-p}$ is the F-distribution with p and m-p degrees of freedom at significance level α. This limit is obtained following the commonly used upper control limit of the Hotelling's T^2 control chart. Furthermore, the contribution of each feature to the T^2 statistic can also be calculated using:

$$d_j = T^2 - T^2_{(j)} \tag{6}$$

where, T^2 is the current value of the statistic while $T^2_{(j)}$ is the value of the statistic using all the features except for the j^{th} feature.

One potential problem is that there may not be enough data at $S_t^{(c)}$ to estimate the mean vector and covariance matrix in Equation (4). However, more data from each category is expected to be collected in the future and also if there are not enough data points for computing the Hotelling T^2 statistic then this means that not many worn-out moulds passed the $S_t^{(c)}$ number of shots.

3. Data analysis

Data processing

The first part of the analysis focuses on the cluster analysis and for this, the mould characteristics and pre-chosen running settings data sets are used. The data sets contained an abundance of missing values (NAs). Based on our discussions with the industrial partner, it was concluded that the missing observations were due to the specific characteristics of a mould and not due to measurement errors. For this, it was decided to encode the columns with NAs using a binary categorical variable with the categories, "missing" and "recorded".

The next step of the analysis involves the production and maintenance data sets. Here, it was necessary to combine two data sets originating from different sources and recorded at different frequencies. In order to overcome this challenge, the average over the time intervals is computed which will ensure the same sampling frequency for all features.

Clustering results

When the data sets on mould characteristics and pre-chosen running settings are merged, there are roughly 900 observations and more than 200 features. Using the unsupervised random forest model, the variable importance can be computed in terms of mean decrease Gini index. The first thirty features are presented in Figure 4 with "Mould Characteristic or Setting 182 #" being the most important feature.

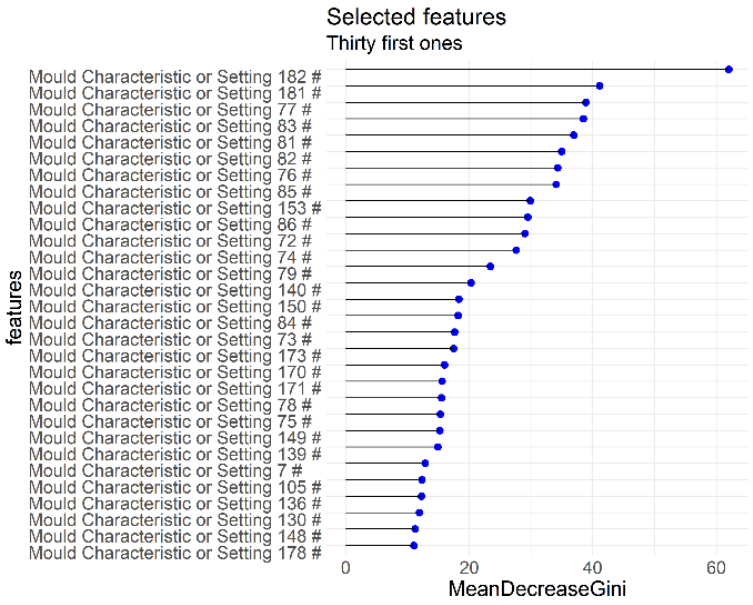

Figure 4. First thirty features given by the mean decrease in the Gini index.

The dissimilarity matrix for the moulds is obtained from the unsupervised random forest model. In Figure 5, using the dissimilarity matrix, the average silhouette width based on PAM is computed along with the t-SNE clustering using four clusters.

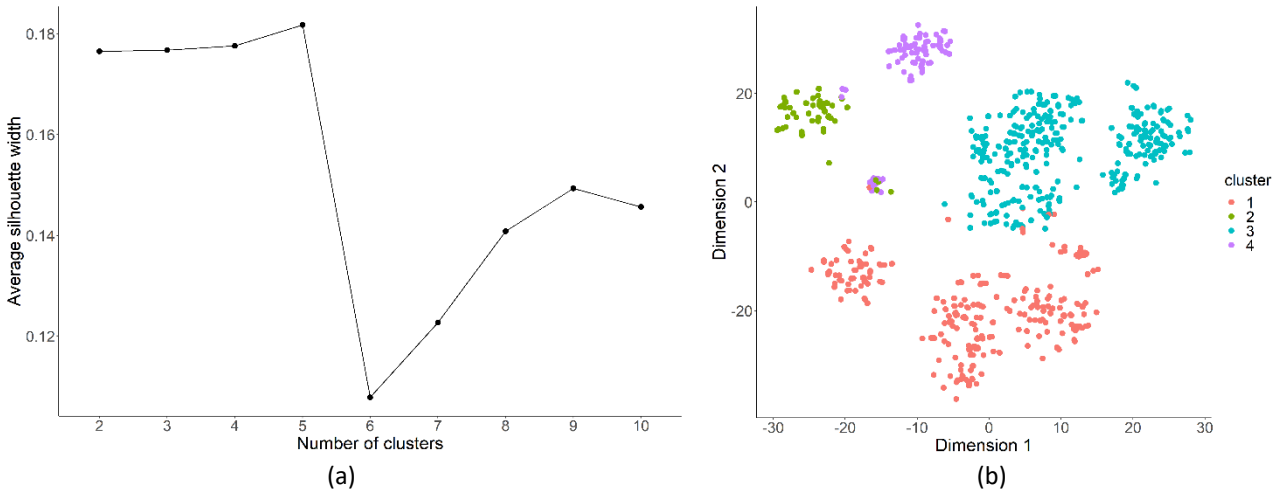

Figure 5. (a) Average silhouette width and (b) t-SNE clustering using four clusters.

Based on the average silhouette width in Figure 5a, it seems that the observations can be grouped in 5 clusters. Since the number of clusters of 2, 3 and 4 have average silhouette values close to the case with 5 clusters, t-SNE plots are also consulted and it is found that 4 clusters provide the best clustering of the data.

To simplify the interpretation of the clustering algorithm, solely the “Mould Characteristic or Setting 182 #” feature is considered as it has the highest importance value in the random forest model. In fact, this feature is further confirmed by the industrial partner as an important feature in production. As shown in Figure 6a, this feature alone separates the data very well with the exception of certain categories. Therefore, due to the practical interpretation and simplicity, for the next step of the analysis, the clustering given by the “Mould Characteristic or Setting 182 #” feature with the five categories is used as in Figure 6b.

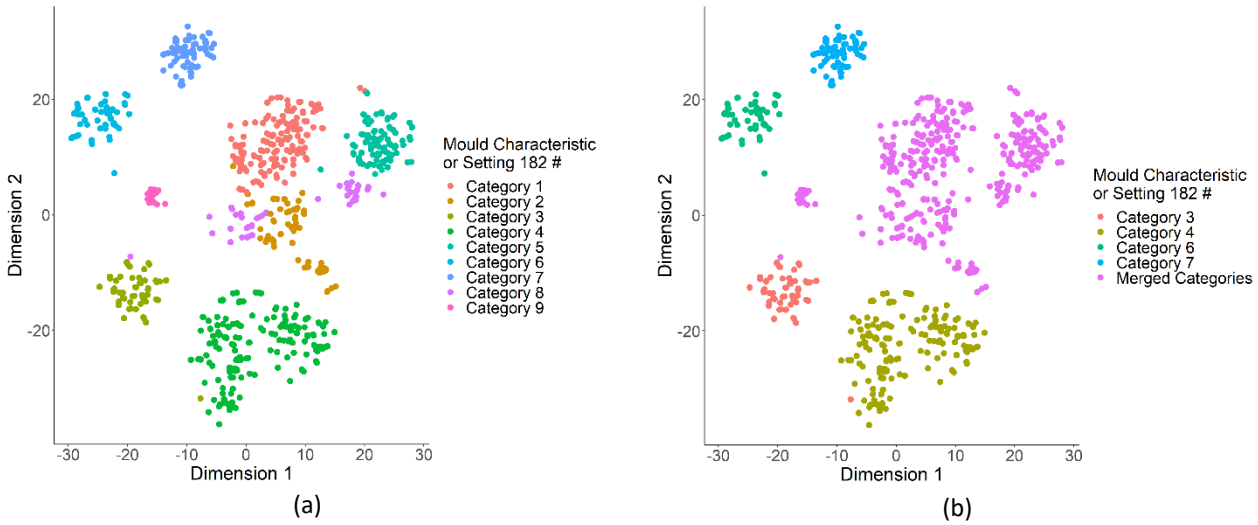

Figure 6. (a) t-SNE clustering using all categories and (b) t-SNE clustering using five selected categories.

Survival curves using the Kaplan-Meier estimator and random survival forest

In the available data set, only 10 % of the moulds are worn-out while the rest are still in use. Initially, the survival curves are estimated using the Kaplan-Meier estimator. Figure 7a shows a difference between the survival curves from different categories. The same conclusion is also reached after running the log rank test at a significance level of $\alpha = 0.05$. In Figure 7b, after also including the running moulds data, there are small differences in the survival curves that were based only on the worn-out data. It can be observed that more data is needed for the "Merged Categories" and "Category 7" for getting a better estimate of the survival curves.

The most important conclusion based on Figure 7 is that the clustering categories are beneficial for getting an overview of the mould's lifetime and can be used in production for getting an overview of the mould's state during production.

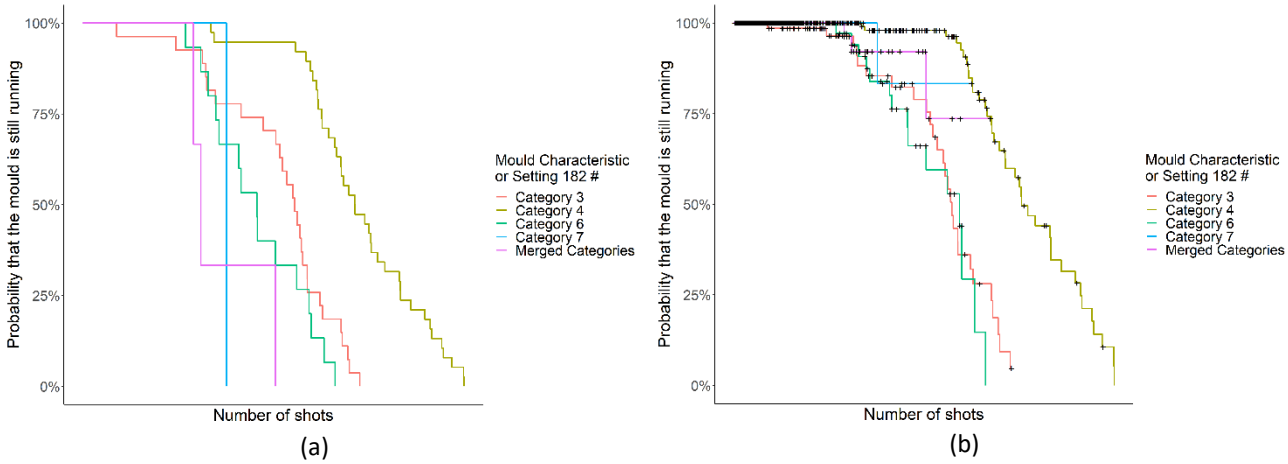

Figure 7. (a) Kaplan-Meier survival curve estimates using only worn-out data and (b) Kaplan-Meier survival curve estimates using worn-out and running data (censored data is marked with crosses).

What is important to mention at this stage is that during part production, a mould can be taken out for repair based on certain information that is unfortunately missing in the current case study. One example is given by the visual inspection of product quality, which is a good indicator when a mould is worn-out. However, this data set will only be available in the future, which means that a model using this information is plausible and needs to be considered. At present, the production and maintenance data sets are available

and thus, only this data is incorporated into a random survival forest model, mostly for illustration purposes. The main point of interest is to compare the random survival forest model with the Kaplan-Meier estimated survival curves. A number of 10000 trees are used for fitting the random forest model and the worn-out data for training. After the model was trained, an OOB c-index of 0.85 was obtained, which indicates that the model is performing better than random. In Figure 8, the Kaplan-Meier survival curves are compared with the random survival forest survival curves from the worn-out data.

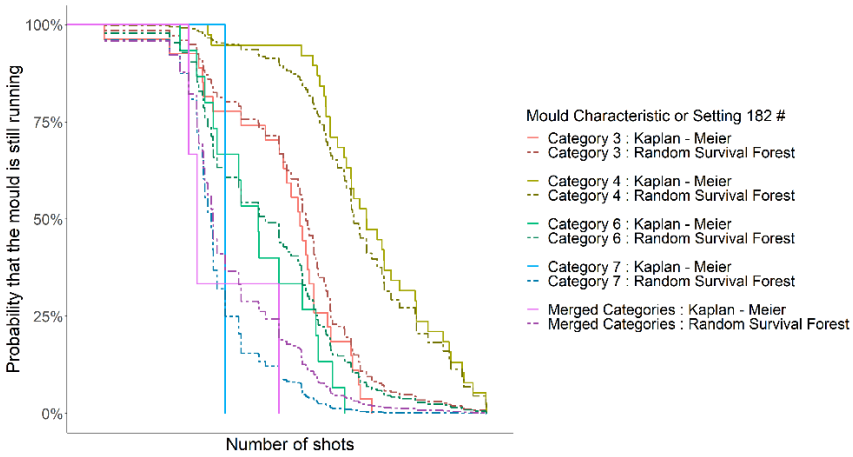

Figure 8. Comparison between Kaplan-Meier survival curve estimates and random survival forest on worn-out data.

Figure 8 shows that for “Category 3” and “Category 4”, the curves are behaving similarly with a small difference towards the end of the life of “Category 3”. For other categories however, the differences are more pronounced. On the other hand, the data in these categories is taking full advantage of the random survival forest model. Therefore, it was decided as a short term solution to use the Kaplan-Meier estimate of the survival curve and for the future, when more data is available, to use a random survival forest model. For a practical implementation in production, the authors are proposing a solution in the form of a dashboard using R Shiny (2019) which is described in detail in the next section.

4. R Shiny dashboard

The results presented above are incorporated into a dashboard such that it is easy for the industrial partner to monitor the status of the moulds. The R Shiny implementation is advantageous since it can easily be used on cloud or other data management platforms. The dashboard focuses on two different aspects. Firstly, the Kaplan-Meier survival curves are presented based on the historical data as well as using the running moulds data. Secondly, a way to monitor the running moulds in terms of conditional survival curves based on Kaplan-Meier estimates is presented. Furthermore, the early warning system for a running mould is also presented.

Historical data

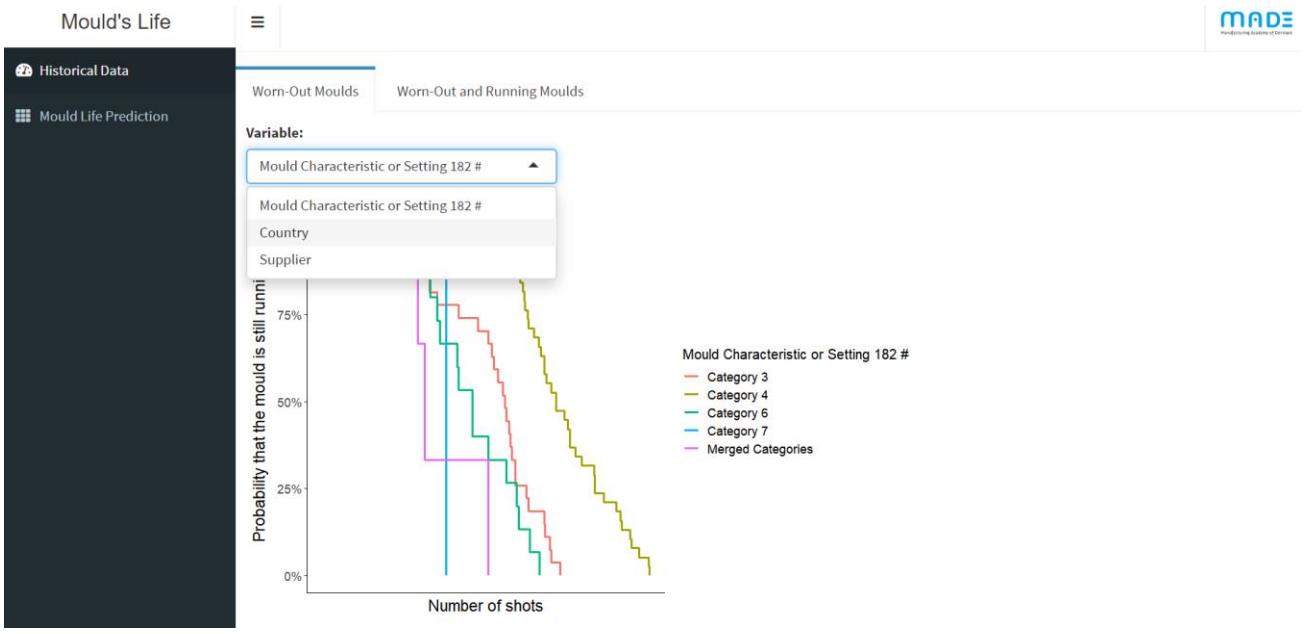

Figure 9. Snapshot from the Kaplan-Meier survival curve estimates for worn-out data.

In Figure 9, on the left panel in black, there are two options as a choice, namely “Historical Data” and “Mould Life Prediction”. The “Historical Data” option is highlighted which further contains two tabs, i.e. “Worn-Out Moulds” and “Worn-Out and Running Moulds”. The “Worn-Out and Running Moulds” contains the exact same plot as “Worn-Out Moulds” with the exception that the survival curves are built using all the available data, i.e., worn-out and running moulds data. The “Worn-Out Moulds” tab contains only the survival curves built on the worn-out moulds data. The “Variable:” field contains three different choices. The first choice is given by the five categories of “Mould Characteristic or Setting 182 #” which provides a quick overview over the historical worn-out moulds. Furthermore, since the industrial partner has different production sites, it was decided to provide survival curves computed using the Kaplan-Meier estimate for different countries. The same is valid for different mould suppliers. In this way, the industrial partner can have a convenient global overview over the worn-out moulds. Crucially, the curves can be updated in real-time. Thus, each time a new worn-out mould is stored in the database, the survival curves estimates can be updated instantaneously. Deciding when to update the database is also an important question but it is beyond the scope of this study. The option to also view the survival curves using the running moulds data was included so that the industrial partner can check if the running moulds are changing in some way the behavior of the survival curves in different regions.

Mould life prediction

In Figure 10, the “Mould Life Prediction” choice is highlighted. There are two tabs associated with this choice. The first tab “Using Worn-Out Moulds Data”, contains the conditional survival curve (with the confidence interval) based on the category of the running mould using only the worn-out data. Moreover, this tab also contains the early warning system for the running mould. The second tab, “Using Worn-Out and Running Moulds Data” contains the conditional survival curve based on all the data which means worn-out and running data without the early warning system.

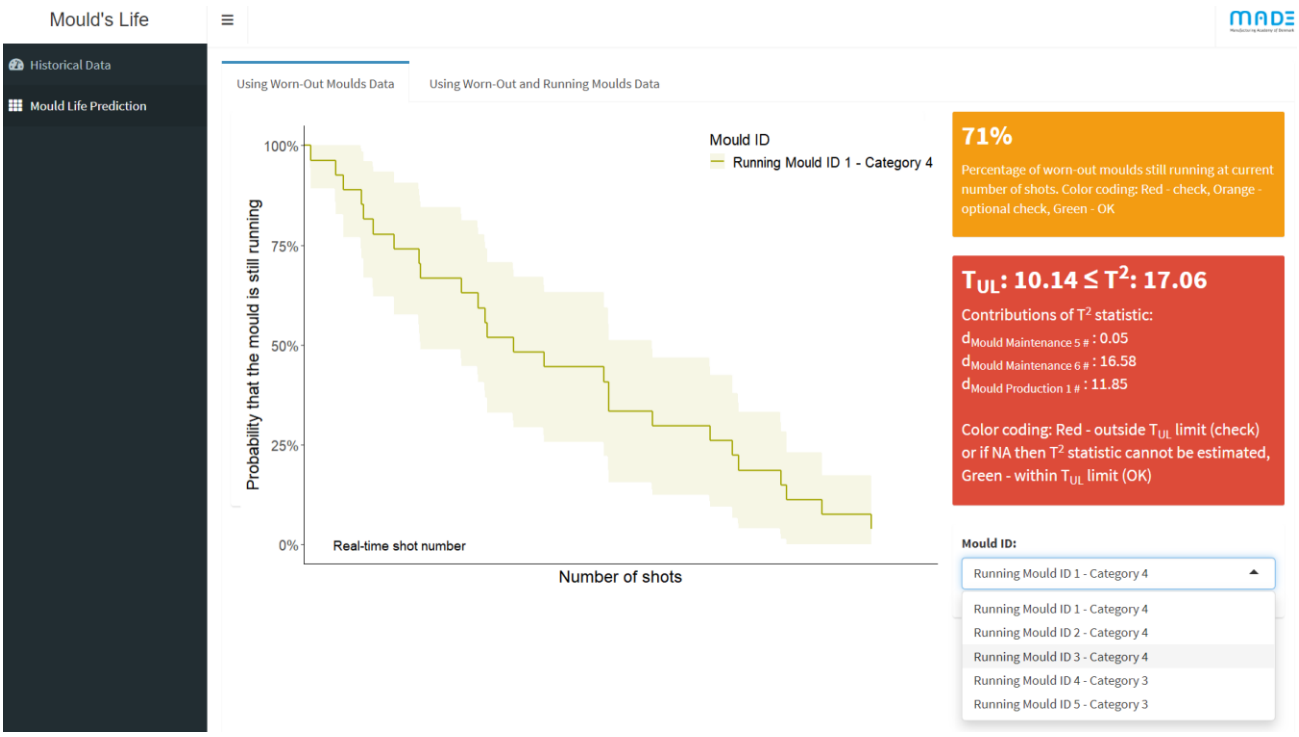

Figure 10. Snapshot from the early warning system along with the conditional survival curve for a running mould based on the Kaplan-Meier estimate.

In Figure 10, the conditional survival curve can be visualized along with the 95 % confidence interval for a running mould in "Category 4" at the corresponding running shot number. Furthermore, on the right side, two different boxes are presented. The first box presents the percentage of the active moulds from the worn-out moulds at that shot number. Here, there are 71.1 % of the worn-out moulds still running at this number of shots. It was decided to color the boxes according to the "risk" associated with the percentage of available historical worn-out moulds. For example, if the percentage lies between 0 and 20 % then a red color is used, whereas if the percentage lies between 20 and 80 % then the color becomes orange. If the percentage lies between 80 and 100 % then the box is colored green. These categories were chosen only for illustration purposes. In the industrial setting, these values should be adjusted by the operators' needs. Using the color coding, the industrial partner has an immediate overview of the status. The second box (in red) presents the Hotelling T^2 statistic along with the T_{UL} limit. Also, this box is color coded dynamically based on the results. If the box is green, the Hotelling T^2 statistic is within the limit whereas if the box is red, then the statistic lies outside the limit or the statistic cannot be estimated. In addition, the contributions for three features are also presented. These features were chosen based on some discussions with the industrial partner. The early warning system, can immediately point out if the running mould is behaving similarly in terms of production and maintenance with the historical data as the statistic is built on the average at that number of shots of the worn-out mould data.

As an upgrade for the "Mould Life Prediction" choice, one could use a random forest model instead of the Kaplan-Meier survival curves estimates, thus, incorporating also other production data. In terms of the early warning system, one could use a combined Hotelling T^2 statistic along with a Q chart with the corresponding contributions.

5. Conclusion and future outlook

In this article, a case study using industrial data from the plastic injection moulding industry is presented. The problem considered in the case study is the prediction of worn-out moulds with a focus on visualization and presentation in real-time through the means of an R Shiny dashboard. The analysis is divided into two

parts. First, the authors focus on clustering using mixed data and then on worn-out mould prediction. Using the results of the cluster analysis, the worn-out moulds are presented in terms of survival curves using the Kaplan-Meier estimator. The cluster analysis shows that one feature in particular can characterize the moulds very well. Moreover, a comparison between the survival curves using the Kaplan-Meier estimate and a random survival forest model is presented.

In terms of real-time monitoring of the running moulds, a R Shiny dashboard has been implemented. One tab of the dashboard shows the survival curves estimates via the Kaplan-Meier estimator using both worn-out moulds as well as all available data including the data on the running moulds. The other tab focuses on presenting the conditional survival curve estimates for a running mould including an early warning system. This system is based on a percentage that indicates the proportion of the running worn-out moulds at that number of shots and a Hotelling T^2 statistics along with the production or maintenance feature contributions. To the best of our knowledge the early warning system has not been implemented for this type of problem before and also not in this manner.

As future work, the aim is to incorporate all available data into a random forest model for getting a better survival curve prediction. The data in this case study was unfortunately missing product metrology data, which it is considered important for assessing the mould wear and tear. Other matters that will be interesting to investigate are data management related issues. For example, how often should the training data used for model building be updated, and how much data should be included in the training data? In terms of the early warning system, if the p features are highly correlated, then the Hotelling T^2 and Q statistics (De Ketelaere, Hubert, and Schmitt 2015) using Principal Components Analysis (PCA) (Hastie, Tibshirani, and Friedman 2013) 2013) can be used. The contributions can also be computed for this case as presented in the work of Miller, Swanson, and Heckler (1998) and De Ketelaere, Hubert, and Schmitt (2015). Moreover, a combined statistic based on Hotelling T^2 and Q statistics (e.g. Frumosu and Kulahci 2019) could be used instead of just a Hotelling T^2 statistic such that more features could be used for monitoring along with the corresponding contributions presented as a bar chart.

Acknowledgments

This work has been carried out under MADE SPIR – Strategic Platform for Innovation and Research, Denmark and the authors are grateful for the given opportunity. The authors would like to thank Max Peter Spooner for proofreading the article and the industrial partner for providing the data sets and for the allocated time.

Disclosure statement

The authors of the article have not encountered any potential conflict of interest.

References

Ahmad, A., and S.S. Khan. 2019. “Survey of State-of-the-Art Mixed Data Clustering Algorithms.” Ieee Access 7 (99): 31883 – 31902. doi:10.1109/ACCESS.2019.2903568.

Ahmad, R. and S. Kamaruddin. 2012. “An overview of time-based and condition-based maintenance in industrial application.” Computers and Industrial Engineering 63 (1): 135-149. doi:10.1016/j.cie.2012.02.002.

Aivaliotis, P., K. Georgoulias, and G. Chryssolouris. 2019. “The use of Digital Twin for predictive maintenance in manufacturing.” International Journal of Computer Integrated Manufacturing 32(11): 1067-1080. doi: 10.1080/0951192X.2019.1686173.

Alsina, E., M. Chica, K. Trawiński, and A. Regattieri. 2018. “On the use of machine learning methods to predict component reliability from data-driven industrial case studies.” International Journal of Advanced Manufacturing Technology 94 (5-8): 2419-2433. doi:10.1007/s00170-017-1039-x.

Amari, S. V., L. McLaughlin, and H. Pham. 2006. “Cost-effective condition-based maintenance using Markov decision processes.” In RAMS'06 Annual Reliability and Maintainability Symposium, 464-469. IEEE. doi:10.1109/RAMS.2006.1677417

Boström, H., L. Asker, R. Gurung, I. Karlsson, T. Lindgren, P. Papapetrou. 2018. “Conformal prediction using random survival forests.” Proceedings - 16th Ieee International Conference on Machine Learning and Applications, Icmla 2017 2018: 812-817. doi:10.1109/ICMLA.2017.00-57.

Breiman L. 2001. “Random forests.” Machine Learning 45 (1):5-32. doi:10.1023/A:1010933404324.

Breiman , L., and A. Cutler. 2019. Random Forests. University of California, Berkley. Accessed June 21 2020. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Bukkapatnam, S.T.S., K. Afrin, D. Dave, and S.R.T. Kumara. 2019. ”Machine learning and AI for long-term fault prognosis in complex manufacturing systems.” Cirp Annals 68 (1): 459-462. doi:10.1016/j.cirp.2019.04.104.

Cachada, A., J. Barbosa, P. Leitão, C. A. S. Geraldes, L. Deusdado, J. Costa, C. Teixeira, J. Teixeira, A. H. J. Moreira, P. M. Moreira, and L. Romero. 2018. “Maintenance 4.0: Intelligent and predictive maintenance system architecture.” In 2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA), 139-146. IEEE. doi:10.1109/ETFA.2018.8502489.

Da Silva, V. L., J.L. Kovaleski, R.N. Pagani, J.D.M. Silva, and A. Corsi. 2020. “Implementation of Industry 4.0 concept in companies: Empirical evidences.” International Journal of Computer Integrated Manufacturing 33 (4): 325-342. doi: 10.1080/0951192X.2019.1699258.

De Ketelaere, B., M. Hubert, and E. Schmitt. 2015. “Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data.” Journal of Quality Technology 47 (4): 318-335. doi:10.1080/00224065.2015.11918137.

Foss, A. H., M. Markatou, and B. Ray. 2019. “Distance metrics and clustering methods for mixed-type Data.” International Statistical Review 87 (1): 80-109. doi:10.1111/insr.12274.

Frumosu, F. D., and M. Kulahci. 2019. “Outliers detection using an iterative strategy for semi-supervised learning.” Quality and Reliability Engineering International 35 (5): 1408–1423. doi:10.1002/qre.2522.

Gower, J. C. 1971. “A General Coefficient of Similarity and Some of Its Properties.” Biometrics 27 (4): 857-871. doi:10.2307/2528823.

Harrell, F. E., K.L. Lee, and D.B. Mark. 1996. “Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.” Statistics in Medicine 15 (4): 361-387.doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Harrell, F.E., R. Califf, D. Pryor, K. Lee, and R. Rosati. 1982. “Evaluating the yield of medical tests.” Journal of the American Medical Association 247: 2543–2546. doi:10.1001/jama.1982.03320430047030.

Harrington, D. P., and T.R. Fleming. 1982. “A class of rank test procedures for censored survival data.” Biometrika 69 (3): 553-566. doi:10.1093/biomet/69.3.553.

Hashemian, H. M. 2010. “State-of-the-art predictive maintenance techniques.” IEEE Transactions on Instrumentation and measurement 60 (1): 226-236. doi:10.1109/TIM.2010.2047662.

Hastie, T., R. Tibshirani, and J. Friedman. 2013. The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

Hieke, S., M. Kleber, C. König, M. Engelhardt, and M. Schumacher. 2015. “Conditional survival: A useful concept to provide information on how prognosis evolves over time.” Clinical Cancer Research 21 (7): 1530-1536. doi:10.1158/1078-0432.CCR-14-2154.

Haarman, M., M. Mulders, and C. Vassiliadis. 2017. Predictive Maintenance 4.0 - Predict the unpredictable. PwC and Mainnovation. Accessed June 28 2020. https://www.pwc.nl/nl/assets/documents/pwc-predictive-maintenance-4-0.pdf

Ishwaran, H., U. Kogalur, X. Chen, and A. J. Minn. 2011. "Random survival forests for high-dimensional data. Statistical Analysis and Data Mining." 4 (1): 115-132. doi:10.1002/sam.10103.

Ishwaran, H., U.B. Kogalur, E.H. Blackstone and M.S. Lauer. 2008. "Random survival forests." Annals of Applied Statistics 2 (3): 841-860. doi:10.1214/08-AOAS169.

Jardine, A.K.S., D. Lin, and D. Banjevic. 2006. "A review on machinery diagnostics and prognostics implementing condition-based maintenance." Mechanical Systems and Signal Processing 20: 1483–1510. doi:10.1016/j.ymssp.2005.09.012.

Kaplan, E. L., and P. Meier. 1958. "Nonparametric Estimation from Incomplete Observations." Journal of the American Statistical Association 53 (282): 457-481. doi:10.1080/01621459.1958.10501452.

Kaufman, L., and P.J. Rousseeuw. 1987. "Clustering by means of medoids." Statistical Data Analysis Based on the L1-norm and Related Methods. First International Conference. 405-16.

Lasi, H., H-G. Kemper, P. Fetteke and T. Feld. 2014. "Industry 4.0." Business and Information Systems Engineering 6: 239–242. doi:10.1007/s12599-014-0334-4.

Li, Z., K. Wang, and Y. He. 2016. "Industry 4.0 - Potentials for predictive maintenance." Proceedings of the 6th International Workshop of Advanced Manufacturing and Automation, 42-46, Atlantis Press.

Li, Z., Y. Wang, and K.S. Wang. 2017. "Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario." Advances in Manufacturing 5(4): 377-387. doi: 10.1007/s40436-017-0203-8.

Miller, P., R. E. Swanson, and C. E. Heckler. 1998. "Contribution plots: a missing link in multivariate quality control." Applied Mathematics and Computer Science 8 (4): 775-792.

Montgomery, D. C. 2012. Introduction to Statistical Quality Control. 7th Edition. Wiley

Peng, Y., M. Dong, and M.J. Zuo. 2010. "Current status of machine prognostics in condition-based maintenance: A review." International Journal of Advanced Manufacturing Technology 50 (1-4): 297-313. doi:10.1007/s00170-009-2482-0.

R Shiny. 2019. Shiny: Easy web applications in R. RStudio Inc. Accessed 21 June 2020. http://shiny.rstudio.com/

Ragab, A., M. S. Ouali, S. Yacout, and H. Osman. 2016. "Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan - Meier estimation." Journal of Intelligent Manufacturing 27 (5): 943-958. doi: 10.1007/s10845-014-0926-3.

Rousseeuw, P. J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of Computational and Applied Mathematics 20: 53-65.

Sai, V. C., M.V. Shcherbakov, and V.P. Tran. 2019. "Data-driven framework for predictive maintenance in Industry 4.0 concept." In Conference on Creativity in Intelligent Technologies and Data Science, 344-358. Springer, Cham. doi:10.1007/978-3-030-29743-5_28.

Shi, T., and S. Horvath. 2006. "Unsupervised learning with random forest predictors." Journal of Computational and Graphical Statistics 15 (1): 118-138. doi:10.1198/106186006X94072.

Si, X. S., W. Wang, C. H. Hu, and D. H. Zhou. 2011. "Remaining useful life estimation - A review on the statistical data driven approaches." European Journal of Operational Research 213 (1): 1-14. doi:10.1016/j.ejor.2010.11.018.

Van Der Maaten, L., and G. Hinton. 2008. "Visualizing data using t-SNE." Journal of Machine Learning Research 9: 2579-2625.

Wang, P., Y. Li, and C.K. Reddy. 2019. “Machine learning for survival analysis: A survey.” Acm Computing Surveys 51 (6): 110. doi:10.1145/3214306.

Zabor, E. C., M. Gonen, P.B. Chapman, and K. S. Panageas. 2013. “Dynamic Prognostication Using Conditional Survival Estimates.” Cancer 119 (20): 3589-3592. doi:10.1002/cncr.28273.