Excellence In Services International Conference EISIC

Formerly Toulon-Verona Conference, founded in 1998

26 EISIC – 2023

PREDICTION OF TOTAL QUALITY PERFORMANCE

David Runosson
Department of Logistics and Quality Management
Linköping University, Sweden
david.runosson@liu.se
and
Holmen AB, Stockholm, Sweden

Peter Cronemyr
Department of Logistics and Quality Management
Linköping University, Sweden
peter.cronemyr@liu.se

Abstract

Background and purpose of the paper: In the context of the fourth industrial revolution, the possibility of developing prediction models has garnered attention for their ability to anticipate product properties based on process settings. In paperboard making, in which you have low sampling rates, with time delays to sampling and about 2000 process variables that affect the outcome, it is easy to believe that these types of prediction models drastically can improve process yield by providing guidance to operators. However, paperboard making is evaluated by multiple quality properties that intercorrelate with each other and the deployment of numerous prediction models without understanding of normal process variation can lead to an information overload, causing tampering and lead to a reduced quality output.

Methodology: This paper investigates the feasibility of using machine learning models to predict total quality yield on a full-scale paperboard machine.

Main Findings: By changing the focus to an aggregated level, our study aims to alleviate the burden of information overload and reduce the risk of tampering.

Originality/Practical implications: The research shows that it is possible to predict the total quality yield with reasonable accuracy. The approach seems promising, and it is believed that it could provide insight for operators and support operational management.

Type of paper: Full scale case study

Keywords
Prediction; Machine learning; Tampering

1. Background – too much data but too slow quality control

Process Management, traditionally characterised as ‘control and stability’, can suffer from slow improvements (Cronemyr and Smeds, 2022). The emergence of Big Data Analytics (BDA) has made fast improvements possible. However, with BDA, there is a risk of getting ‘fooled by randomness’ (Taleb, 2005), and the faster the pace, the bigger the risk of tampering and thus, increasing variation and, hence decreasing quality (Cronemyr and Elg, 2014). Although tampering is an old concept, it is under-researched and in need of further investigation (Smeds, 2022). Cronemyr and Smeds (2022) proposed to further investigate several questions within this topic, among them:

- How can we make more actions that are ‘*quick and clean*’ (without tampering) instead of ‘*quick and dirty*’ (with tampering) or ‘*slow and clean*’ (mostly without tampering but too slow)?
- How could tampering be avoided by interpreting data correctly and making the correct actions?

In this paper we investigate these two questions in the context of paperboard making. The fourth industrial revolution has given possibility of developing prediction models that have garnered attention for their ability to anticipate product properties based on process settings. In paperboard making, in which you have low sampling rates, with time delays to sampling and, in this specific case, about 2000 process variables (sensors and setpoints) that affect the outcome, it is easy to believe that these types of prediction models drastically can improve process yield by providing guidance to operators. However, paperboard making is evaluated by multiple quality properties that intercorrelate with each other and the deployment of numerous prediction models without understanding of normal process variation can lead to an information overload, causing tampering and lead to a reduced quality output.

1.1 Purpose of the paper

The purpose of this paper is to elaborate on a partly novel approach to controlling quality output by analysis and reduction of many process variables, hopefully leading to ‘*quick and clean*’ quality control of paperboard making processes.

1.2 Research questions

- ***RQ1:*** Is it technically possible to continually predict The Total Quality Yield on a paperboard machine based on the process variables?
- ***RQ2:*** What are potential benefits of applying Statistical Process Control on Key Performance Indicator level in paperboard making?

1.3 Contributions

David Runosson, an industrial Ph.D. student at Holmen AB and Linköping University, was the main contributor to the conducted research and the paper. Peter Cronemyr, senior associate professor at Linköping University, has supervised the research and provided some input on the background and the theoretical frame of reference, as well as some discussion and conclusions.

2. Methodology

This paper merges two research paths. One path follows the work of Cronemyr and Elg (2014) and Cronemyr and Smeds (2022), who argued that combining fact-based decision making and BDA can help decision makers not to take action on random data, thereby avoid tampering. The second path follows the work of Runosson and Skoglund (2022) and investigates the feasibility of combining the traditional principles of Statistical Quality Control with algorithms used in the field of machine learning.

3. Theoretical background

Here we present the theoretical frame of reference.

3.1 Big Data Analytics and fact-based decision making

Birch-Jensen *et al.* (2020) have highlighted the ‘increasing clock-speed of processes’ which is the speed of data; both internal process data and external customer feedback data. They argue that while customers require faster and faster responses to requests and problems, the clock-speeds of quality and process improvements are still slow in organisations. They argue that “*managers must be able to both address quick improvements through channelling and processing as well as work with more long-term knowledge creation*” (Birch-Jensen *et al.*, 2020, p.824). Here we see a problem caused by too high speed; actions are taken *deterministically* but there is no time to analyse data *probabilistically* to find *real* root causes and then update procedures to avoid recurring problems.

Fact-based decision making is a key component in Quality Management. One of the so-called corner stones of Total Quality Management – TQM – is ‘base decisions on facts’ (Bergman and Klefsjö, 2010). It involves information and analysis of data for the purpose of maintaining customer focus, to drive quality improvement and enhance performance. This is carried out by collecting and analysing information on for instance customer needs, organizational problems, and improvement initiatives. The process from defining specific problems, choosing data to collect, analysing data, and improving performance is supported by a large number of methods and techniques. A well-known and well-established general methodology that encompasses this is Six Sigma (Bergman and Klefsjö, 2010; Cronemyr and Elg, 2014).

Tampering may be an old concept that is well known among quality management practitioners and researchers, but it is still under-researched (Smeds, 2022). The traditional view of tampering builds on Shewhart’s (1931) ideas of the need to distinguish between and actions taken based on common and special causes of variation. Making a type I error, namely taking action to eliminate common causes of variation as they were special causes of variation is a common description of tampering (Smeds, 2022). Thus, tampering is typically mentioned in connection to data analysis aided by Statistical Process Control (SPC) using control charts.

In SPC, a type I error means reacting on a single value *within* control limits, as if it were a special cause, and a type II error means reacting on a single value *outside* of control limits with a major ‘improvement’ of the whole process. Both errors inadvertently lead to increased variation. In this context, synonyms of tampering such as overcontrol and overadjustment are sometimes used.

Another benefit of SPC is to indicate what numerical goals of a process that can be considered realistic. Setting a goal for the performance beyond control limits is “*as sensible as to defy gravity*” (Deming, 1993, p. 42), see Figure 1. When managers try to control the process by setting unrealistic targets, it is yet another type of tampering that could be avoided by using SPC. “*What we need is a method for improvement of the process*” (Deming, 1993, p. 43). That method is to first measure, monitor and control the process, using SPC, and then to find and reduce root causes to unwanted variation.

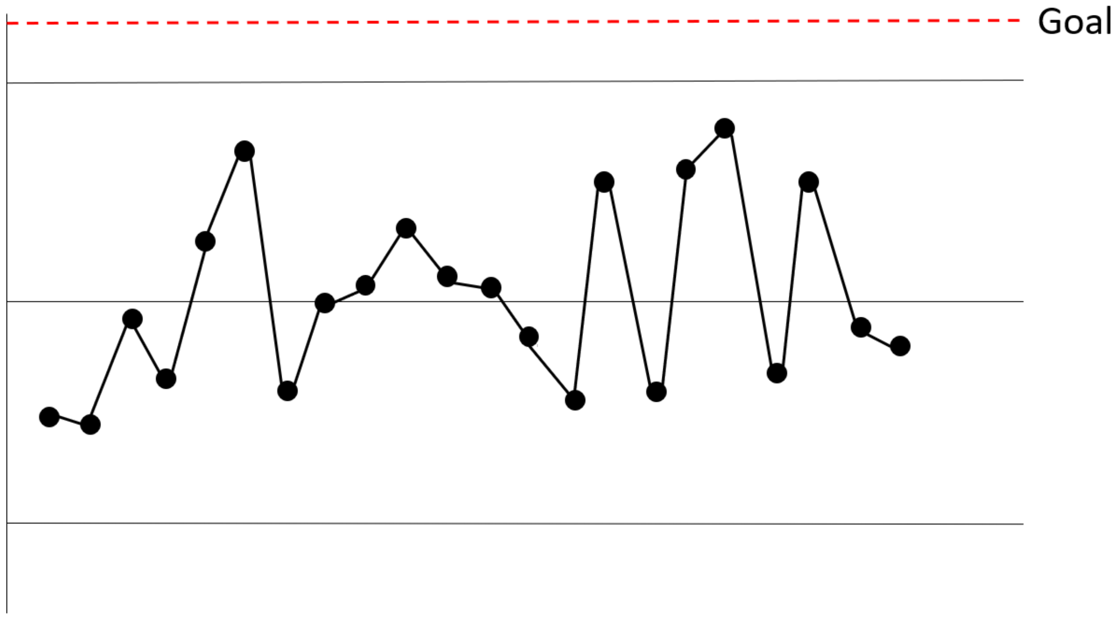

Figure 1: Control chart showing an unrealistic goal of a process in control. (Based on a figure from Deming, 1993, p.42).

3.2 Statistical Process Control, Principal Component Analysis, and Machine Learning

Control charts in SPC serve the purpose of detecting abnormal quality property variations by comparing observed data points to control limits, which are determined based on the statistical behaviour of historical data. Monitoring several quality properties using many univariate control charts simultaneously can however be misleading since it will be discarding all information that considers the joint probability. By using univariate control charts, correlation between quality properties will not be considered and we increase the risk of false negative error. To consider the joint probability one can use multivariate control charts such as the Hotelling T2 chart which enables the one to visualize two quality properties on one single control chart. As the number of variables grow, one can consider dimension reduction using Principal Component Analysis (PCA) (Montgomery, 2009).

Skoglund et al. (2004) showed that by combining the Hotelling T^2 chart with PCA, it is also possible to monitor the process stability of a paperboard machine just by adding process variables instead of quality properties. By monitoring process variables, we can check for stability but not capability. So, Hotelling T^2 chart with PCA is useful for getting a *predictable* process (in process variables) but it does not tell us whether the process is *good*. To decide that a quality output variable is needed. With specification limits a capability analysis can be carried out with the answer expressed as capability indices or The Total Quality Yield.

Machine learning is the name of the subfield of artificial intelligence, which originally was defined as "the field of study that gives computers the ability to learn without explicitly being programmed" by AI pioneer Arthur Samuels (Massachusetts Institute of Technology, 2021). It differs from traditional data analysis in such a way that the analytics is automated by a computer program that learns from previous observations, so-called training data. The computer program in this meaning, will be a mathematical model that describes the relationship between variables that correspond to properties of interest (Lindholm et al., 2022).

There are different fields within machine learning. Unsupervised machine learning, which includes PCA, aims to cluster data in groups without trying to explain the outcome of any specific variable directly. Supervised learning is the field in which the mathematical model aims to predict the result of one or many output variables based on the settings of one or many input variables. This is done with either classification models that predicts a discrete output or regression models that predicts a continuous output (Joshi, 2023).

There are many types of models for both regression and classification suited for different types of problems. It is therefore important to the correct assumptions about the data in order to create a reliable prediction model (Dietrich et al., 2015).

Prediction models has been used for successfully predicting many quality properties in industry e.g. Rodríguez-Álvarez *et al.* (2022) who predicted the basis weight of paper and Pauck *et al.* (2014) who predicted the brightness and deinking behaviour of recycled paper.

The methodology of Random Forest (Jooshi, 2023) uses so called decision trees, which use another type of mathematical approach than many other machine learning models. While many other approaches only handle data in a numerical way, decision trees classify data into subgroups. If the input variable is discrete, the subgroups will be based on the discrete options. If the input variable is continuous, the subgroups will be based on thresholds on that variable, i.e. over or under a certain threshold. Due to this core logic, decision trees will be able to handle both linear and non-linear relations between the input and output variable. Although the algorithm classifies the data, it can also be used for regression. The output will be stepwise based on how many subgroups you allow the algorithm to create. (Jooshi, 2023)

Random Forest uses a set number of regression trees in combination. Each decision tree is built using a different subset of the data, created by both bootstrapping and random feature selection. The bootstrapping is done by randomly creating subsets of data with replacement and thereby creating multiple datasets from the original data. The data that was not used for the training, the so called Out-Of-Bag data, is used to test the model's accuracy. The random feature selection only considers a random subset of variables for each tree. This reduces correlation between trees and thereby reduces the risk of overfitting.

For regression problems, the prediction is created by averaging the outputs from all the decision trees. The number of trees is set by the user. (Lindholm Et al., 2022).

In mathematical terms, the Random Forest regression algorithm is constructed as follows. A Random Forest is created by a number of trees based on a random vector ϴ. We assume that training data is randomly sampled from the distribution of the random vectors Y & X. For any numerical prediction h(x), is calculated accordingly:

$$E_{X,Y}(Y - h(X))^2$$

The Random Forest predictor is formed by averaging over k over the trees $\{h(x, \Theta_k)\}$. As the number of trees in the forest grows, one can with very high probability state:

$$E_{X,Y}(Y - av_k h(X, \Theta_k))^2 \ \rightarrow \ E_{X,Y}(Y - E_\Theta h(X, \Theta_k))^2$$

Denote the right-hand side as PE*(forest) as the generalization of the forest and define the average generalization error of a tree as:

$$PE^*(tree) = \ E_\Theta E_{X,Y}(Y - E_\Theta h(X, \Theta_k))^2$$

Also assume that for all ϴ, EY = E_xh(X, ϴ). Then:

$$PE^*(forest) \leq \bar{\rho} PE^*(tree)$$

Where $\bar{\rho}$ is the weighted correlation between the residuals $Y - h(X, \Theta)$ and $Y - h(X, \Theta')$ where ϴ & ϴ' are independent. (Breiman, 2001)

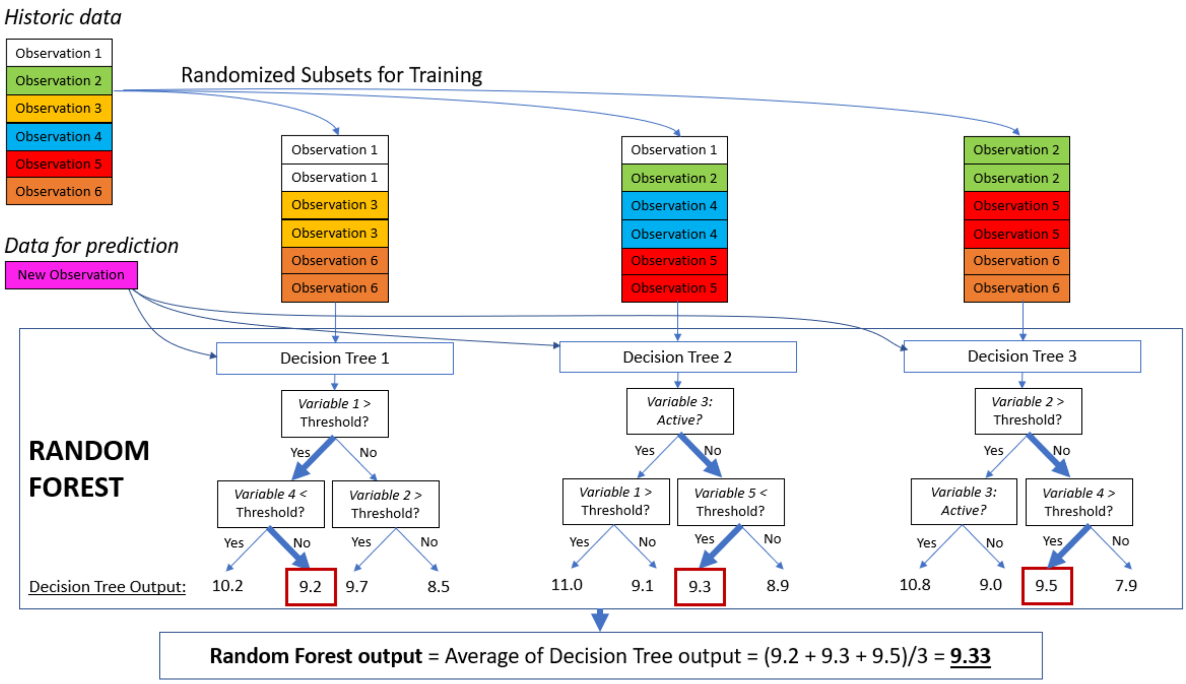

Figure 2: Simplified scheme of the Random Forest algorithm. Indiviudal Decision Tree selections are marked with bold arrows and the individual Decision Tree outputs are marked with a red square.

This means that Random Forest regression is an algorithm that can handle a large number of variables with non-linear relationships. It will also be robust in the sense that overfitting is unlikely, and it can handle outliers, noise, and missing data well since none of the trees in the model can become more dominant than others.

4. Previously used approach

The approach used in this paper is based on a few assumptions:

- The first assumption is that if operational organization detects a problem related to performance target, they wish to act on it.
- The second assumption is that humans are not very good to focus on many targets at the same time.

Paperboard making suffers from a few fundamental problems when it comes to quality and process control. The main one being the low sampling rate for many of the quality properties. Although some properties can be measured online using a traversing measurement head, many quality properties can only be tested in lab. Samples to test are taken by the end of every jumbo reel. The paperboard sample has the full machine width in the cross-machine direction (CD) and about 30 cm in machine direction (MD). Different quality properties have different resolutions in CD from the sampled strip, which is relevant since variation occurs both in MD and CD. The specific paperboard machine in this study is circa 280 meters long and 7 meters wide and produces a jumbo reel every 40 minutes, which then is equivalent to the sampling rate in MD. At a speed of 500-600 m/min, this will be equal to 20-30 km of paperboard.

The low sampling frequency is a problem when it comes process control since you cannot see the effects on actions taken before the next sample is taken and lab tests are performed. Hence, one benefit of implementing machine learning models that can predict outcome of the quality properties based on process signals would be a quicker response for actions taken. However, this is also associated with increased risk of tampering. If adjustments are done without considering the common cause variation it will lead to increased process variation and since many of the quality properties also are intercorrelated with each other; improving one of them may lead to worsening another. This is potentially a problem when using the sampling response for process control, but as the response time decreases the risk for tampering increases. Viewing several quality properties separately also includes a risk for under-controlling the process, since the joint probability will not be considered. Considering this, it is possible that, if one tries to control the process based on a number of independent quality property predictions, one could end up in a state in which one has both tampering problems as well as a under controlled process simultaneously. An existing solution to this is to implement the quality property predictions into a Hotelling T2 chart combined with PCA or, even simpler, use the existing method developed by Skoglund et al. (2004) and surveil the process stability by applying the Hotelling T2 chart combined with PCA on the process variables. However, PCA only considers linear relations between the variables which can neither be assumed for quality properties nor process signals.

From a management perspective, a downside with this approach is that the resulting principal components will be very abstract numbers. You will be able to see if the process is in control or not, but the number will not be relatable directly to common Key Performance Indicators (KPI) such as the Total Quality Yield which management often rely on. In this paper, it is therefore investigated if it is possible to predict Total Quality Yield since this would enable direct responsive univariate SPC directly on KPI level.

Since the reels delivered to the customers are significantly smaller than the full-size jumbo reels produced by the paperboard machine, customer reels are cut out from the jumbo reel. Due to the MD and CD variation, 100% of the jumbo real rarely meets specification. The total quality yield, known as the K-value, is logged for every production shift. It is calculated by dividing the net production of tons paperboard that has been approved by quality control, with the gross production. This yield is used as the target (Y variable) in the prediction model.

5. Novel approach

The 2000+ variables that are continuously logged on the paperboard machine all correlate the outcome to the process to some extent. However, far from all the values can be considered to be directly controllable and the vast majority are not fully independent from one another. To reduce model complexity, it has been attempted to make the prediction by only using variables that are considered to be important for process control. The variable selection was conducted with the help of the distributed control system (DCS), that has existing overview pages for every section of the paperboard machine. These included 97 process variables. Additional variables that were added were the mean and CD standard deviation of those properties that are possible to measure with the traversing measuring head. A few of these properties are measured at the very end paperboard machine and are then considered quality properties. However, most of them are intermediate properties e.g., moisture content at various parts of the process.

Since the resolution of the Y value is very low, with a new value every 8th or 12th hour depending on day of the week, in the training stage, resolution of the X variables was also kept low and more specifically to hourly mean. To reduce model complexity only one product type and grammage was chosen. A 12-month production period was selected for historical data.

The novel model was constructed in Python using the Random Forest Regression algorithm from the scikit learn-package. 80% of the data were used for training of the model, and 20% to validate results. The number of decision trees within the random forest was set to 200.

6. Results

The result shows that the model is able to predict 61% of the variability of the K-value. This shows that it is possible to predict the Total Quality Yield with fair accuracy. The model is far from perfect and has a few clear outliers which are circled in red. This may be indications that more process variables need to be added in the data set or potential special causes that are out of scope for the prediction model, e.g. maintenance related equipment failure, problems related to the incoming pulp or other. Another potential reason might be the low resolution from the actual K-values in the training data, which may cause inaccuracy for less common process settings.

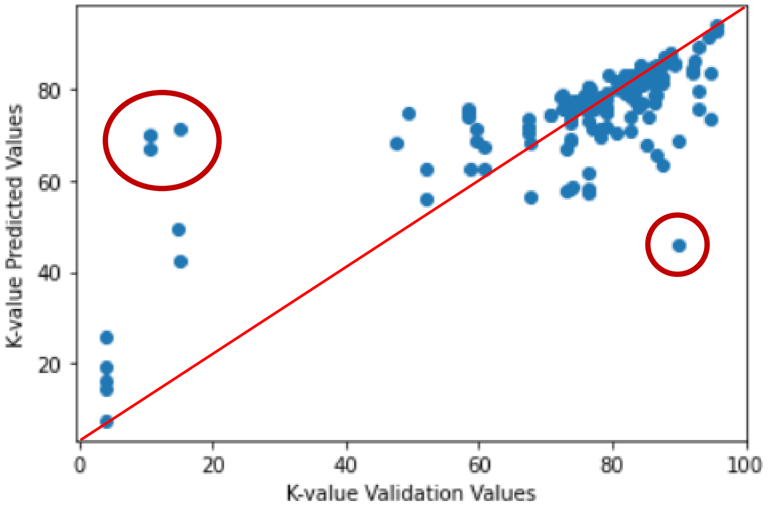

Figure 2: Predicted Values vs Validation Values, r^2=0.61. Outliers circled in red. A 'perfect fit' is indicated by a red line.

To use the model for SPC, these type of causes needs to be addressed. Also, the traditional means to determine the normal state of the process followed by setting the control limits must be conducted.

The results clearly indicate the benefits of using the novel approach over the previously used approach. Once the model has been trained, it can be used on-line, *i.e.* with immediate response for quality output control charts and SPC. The previously used approached could indicate stability but could not evaluate the process quality performance.

7. Conclusions

RQ1: *Is it technically possible to continually predict The Total Quality Yield on a paperboard machine based on the process variables?*

The model created in this paper shows that it is possible to predict the Total Quality Yield with the use of process variables and a Random Forest Algorithm with fair accuracy. It is believed that this model could further be improved with more extensive feature engineering, increased data resolution and removing data that is affected by external factors such as maintenance related equipment failure and other.

RQ2: *What are potential benefits of applying Statistical Process Control on Key Performance Indicator level in paperboard making?*

The potential benefit of using the proposed prediction model combined with Statistical Process Control is that it allows a fast response between process disruptions and actions, while still minimizing the risk of tampering and the risk of neglecting the joint probability interaction of different process parameters. Compared to using existing solutions based on PCA this prediction model would also consider nonlinear interactions and provide an output value that is not only abstract, hence more understandable to operators and managers.

8. Discussion

The model proposed in this paper uses Random Forest Algorithm which is a Black Box Model which means that humans cannot interpret why the model makes a specific prediction.

For an operational organization this is likely to cause frustration if not the cause for a problem promptly can be determined. To solve this problem, one could try alternative prediction algorithms that are interpretable. If a robust prediction cannot be achieved while using an interpretable algorithm, a different approach might be to incorporate algorithms from the field of eXplainable Artificial Intelligence (XAI). These XAI algorithms are developed to help humans to interpret Black Box and how to act if one wants to change the outcome of the model. Further research should also include real testing of prediction on KPI level for process control.

Statistics has for a long time been a cornerstone of quality management. As machine learning i.e., advanced statistics, permeates industries, we believe that quality management should form the guiding principles within this implementation journey. Don't you?

References

Bergman, B. and Klefsjö, B. (2010), *Quality from customer needs to customer satisfaction*, Lund, Studentlitteratur AB.

Birch-Jensen, A., Gremyr, I. and Halldórsson, Á. (2020), Digitally connected services: Improvements through customer-initiated feedback. *European Management Journal*, 38(5), pp. 814-825.

Breiman, L., (2001), Random Forests, Available at: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (Accessed at 2023-06-14).

Brown, S. (2021), *Machine learning, explained.* Available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (Accessed at 2023-06-26).

Cronemyr, P. and Elg, M. (2014), The time is right for Fact-Based Decision Making – Applying QM/QC tools to Big Data. *In the proceedings of Quality Management and Organisational Development International Conference,* Prague, Czech Republic.

Cronemyr, P. and Smeds, M. (2022), Do Big Data Analytics Lead to Tampering? *In the proceedings of EISIC 2022*, Visby, Sweden.

Deming, W.E. (1993), *The new economics for industry, government, education*, Cambridge, MA, MIT Center for Advanced Engineering Study.

Dietrich, D., Heller, B. and Yang, B (2015): *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons. Accessed at: (https://2masteritezproxy.skillport.com/skillportfe/assetSummaryPage.action?assetid=RW$2821:_ss_book:72724#summary/BOOKS/RW$2821:_ss_book:72724 (Accessed:2023-06-14).

Joshi, A.V. (2023), *Machine Learning and Artificial Intelligence*, Cham: Springer Nature Switzerland AG. Available at: (https://link.springer.com/book/10.1007/978-3-031-12282-8) (Accessed: 2023-06-14).

Kahneman, D., Sibony, O. and Sunstein, C.R. (2021), *Noise: a flaw in human judgment,* Glasgow: Harper Collins Publishers.

Lindholm, A., Wahlström N., Lindsten F. and Schön T.B. (2022), *Machine Learning – A First Course for Engineers and Scientists.* Cambridge: Cambridge University Press. Available at (https://smlbook.org/book/sml-book-draft-latest.pdf) (Accessed: 2023-06-14).

Montgomery, D.C. (2009), *Introduction to Statistical Quality Control – A Modern Introduction*. Asia: John Wiley & Sons (Asia).

Pauck J.W., Venditti, R. Pocock, J. and Andrew, J. (2014) Neural network modelling and prediction of the flotation deinking behaviour of recycled paper mixes, *Nordic Pulp & Paper Research Journal,* 29(3), pp. 521-532.

Rodríguez-Álvarez, J.L., López-Herrera, R, Villalón-Turrubiates, I.E., García-Alcaraz, J.L., -Díaz-Reza, J.R., Arce-Valdez, J.L., Aragón-Banderas, O. and Soto-Cabral, A. (2022), Alternative method for determining basis weight in papermaking by using an interactive soft sensor based on an artificial neural network model, *Nordic Pulp & Paper Research Journal*, 37(3), pp. 453-469.

Runosson, D. and Skoglund, A. (2022), Root Cause Analysis in Data Heavy Industry: A proposal for a straight-forward approach. *In the proceedings of EUROMA 2022*, Berlin, Germany.

Shewhart, W.A. (1931), *Economic control of quality of manufactured product*, Milwaukee: ASQ Quality Press.

Smeds, M. (2022), *Exploring Tampering: Towards an Understanding of Why Improvement Efforts Sometimes Fail*. Linköping University, Management and Engineering, Doctoral thesis, urn:nbn:se:liu:diva-183125.

Taleb, N. (2005), *Fooled by randomness: the hidden role of chance in life and in the markets,* New York: Random House.

Joshi, A.V. (2023), Machine Learning and Artificial Intelligence, Cham: Springer Nature Switzerland AG. Available at: (https://link.springer.com/book/10.1007/978-3-031-12282-8) (Accessed: 2023-06-14).