DO BIG DATA ANALYTICS LEAD TO TAMPERING?

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

Magdalena Smeds
Department of Logistics and Quality Management
Linköping University, Sweden
magdalena.smeds@liu.se

Abstract

Purpose of the paper: To elaborate on the idea that the clock-speed of processes controlled with Big Data Analytics (BDA) is increasing to the point that the risk of tampering might become a problem with process instability as a consequence.

Methodology: The paper is based on previous research and the predictions made by the authors.

Main Findings: Traditional Process Management, with ‘control and stability’, can suffer from slow improvements. Today, fast changes are necessary. The emergence of BDA has made this possible. However, BDA is often descriptive and diagnostic, but seldom predictive and prescribing. On the other hand, that could be achieved by combining BDA with QM/QC tools. Unfortunately, this is seldom executed. Without, there is a risk of getting ‘fooled by randomness’, and the faster the pace, the bigger the risk of tampering and thus, increasing variation, and hence decreasing quality.

Practical implications: An enhanced understanding of the connection between BDA and tampering will lead to better decisions and results.

Originality/value: Although tampering is an old concept, it is under-researched and in need of further investigation.

Type of paper: Research idea

Keywords

Big Data Analytics; Tampering

1. Background – too little or too much control?

Traditional Process Management, with ‘control and stability’, can suffer from slow improvements. Today, fast changes are necessary. The emergence of Big Data Analytics (BDA) has made this possible. However, BDA is often descriptive and diagnostic, but seldom predictive and prescribing. On the other hand, that could be achieved by combining BDA with QM/QC tools (Cronemyr and Elg, 2014). Unfortunately, this is still seldom executed. Without, there is a risk of getting ‘fooled by randomness’, and the faster the pace, the bigger the risk of tampering and thus, increasing variation, and hence decreasing quality. Although tampering is an old concept, it is under-researched and in need of further investigation (Smeds, 2022).

1.1. Purpose of the paper

The purpose of this paper is to elaborate on the idea that the clock-speed of processes controlled with BDA is increasing to the point that the risk of tampering might become a problem with process instability as a consequence. We are thereby presenting an idea for further research with the aim of enhanced understanding of the connection between BDA and tampering, leading to better decisions and results.

2. Methodology

This paper follows the work of Cronemyr and Elg (2014) who suggest that combining the research streams Fact-Based decision making and BDA could lead to more informed decisions. The same authors highlight that without understanding variation in BDA there is a risk of getting ‘fooled by randomness’ and thus they propose further research on combining BDA and root-cause analysis in order to not get fooled by randomness (Cronemyr and Elg, 2014). This paper turns perspectives by elaborating on the notion that ill-advised use of BDA – by not considering variation – can mislead decision makers to take action on random data *i.e.* tampering.

The paper’s conceptual contribution is to delineate *i.e.* to explore the relationship between the research streams (MacInnis, 2011) on BDA and tampering. This type of conceptual research is appropriate when the aim is to describe a concept or idea, describe why it should be studied and suggest possible further research paths (MacInnis, 2011). Lack of relevant literature when searching for BDA and tampering/variation in Scopus and Web of Science suggests that the relationship between BDA and tampering has – until now – not been described. The discussion presented in this paper is based on the authors previous research on and experience of the two research streams.

3. Theoretical background

Here we present the theoretical frame of reference.

3.1. The increasing clock-speed of processes

A decade ago, Big Data was a common buzz word and has now become a natural part of modern digitalized management. It has now become a blanket term for any collection of large data sets difficult to process using traditional data processing applications. As an example, Google now processes around 2.5 quintillion (10^{18}) bytes of data per day. However, for a specific organization, the amount of data generated and available is of course smaller but still very large. Big Data is even bigger than it used to be.

The question is, are all these data utilized for fact-based decision making in organizations and if so, how? According to LaValle, *et al*. (2011) many management decisions are based on few (or very limited) and old data. There is a need to move BDA from focusing on better IT applications to focusing on core business and operational functions (Davenport *et al.*, 2012; Cronemyr and Elg, 2014). But what has happened since 2014? We argue that the analytical part of BDA is often lacking, while there has been much development of ways to visualize big data. Cronemyr and Elg (2014) argued that tools well-known to quality practitioners and researchers, *i.e.* QM/QC tools, can be used to explore the full potential of Big Data and hence give better decisions in operative and strategic processes. That has partly been done, see *e.g.* contributions from Chowdhury and Sandén (2015), Sehlin *et al*. (2019), Magnusson and Nordlund (2020) and Bergmark and Linderstam (2021).

So, BDA has become *bigger*, but has it become *better*? As seen by the referred examples above, in many respects yes, but more data and more powerful computers and software are still doing the same thing, just more and faster. This leads to another risk, the risk of over-controlling, or tampering. There is a need of balancing speed and control.

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. One of their respondents said, *"IT works with seconds, whilst R&D talk in terms of years – naturally challenges arise"* (Birch-Jensen *et al.*, 2020, p.819). They argue that "*managers must be able to both address quick improvements through channeling 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 analyze data *probabilistically* to find *real* root causes and then update procedures to avoid recurring problems.

Cronemyr *et al.* (2022) present a framework for Change Driven Process Management (CDPM), how to conduct process improvements that are '*quick and clean*' (*i.e.* correct and fast) instead of '*quick and dirty*' (*i.e.* fast but incorrect) or '*slow and clean*' (correct but too slow). They discuss process management and improvement, but do not directly address the use of BDA.

3.2 Big Data Analytics – BDA

Big Data are characterized by the three Vs (Liu, 2014):

- *Volume*. The amount of data stored in the world is growing exponentially. At the same time, the cost of data storage is dropping in the same pattern.
- *Velocity*. Data are generated, collected, stored and processed with increasing speed to meet the demand of data.
- *Variety*. The sources of big data are everywhere. Databases, documents, emails, *etc*. Everything you do is recorded and stored somewhere.

To deal with these volumes of data, we need to take a probabilistic approach – dealing with many data over time, rather than a deterministic approach – looking at snapshots. Statistical methods play a vital role in quality improvement of manufacturing and service processes. The use of statistical methods has also been developed within the quality field so that not only experts in statistics may address data issues but also engineers, managers, and operators. Cronemyr and Elg (2014) suggested that the growing field of research and conceptual development of Big Data Analytics could be enhanced by incorporating concepts and ideas from Quality Management. Not all researchers of BDA agree though. According to Jordan and Lin (2014) "*Today, there are those who seem to suggest that models are unnecessary, that given sufficient computing power, the relevant patterns will emerge, absent theory*." For example, they refer to Mayer-Schönberger and Cukier (2013) who stated: "*For many purposes, correlation is sufficient and people don't need to know causality*". Cronemyr and Elg (2014) agreed with Jordan and Lin that this is unfortunate and that a statistical, *i.e.* scientific, approach is needed, which is the basis for QM/QC tools. Otherwise, you risk being '*fooled by randomness*' (Taleb, 2004; Kahneman *et al.*, 2021) and react to insignificant patterns. Furthermore, with many data (*i.e.* 'big n') there will always be significant correlations (*i.e.* 'low p') but with little or no practical significance (Jordan and Lin, 2014).

Big Data are used in many different ways, depending on opportunities and needs. One opportunity, based on a very clear need and scientific thinking, is to use Big Data for fact-based decision making. According to McAfee and Brynolfsson (2012) "*Data-driven decisions are better decisions – it's as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition*."

Business Intelligence (BI) systems, *e.g.* Microsoft Power BI and Oracle BI, have been promoted as an effective way to conduct BDA in order to shed light on a wide range of complex issues. It implies the use of data, statistical analysis, explanatory and predictive models to gain insights and act in line with these findings. A key element thus is to identify patterns in data by using a variety of different statistical methods. Predictive models allow for predicting the probability of an outcome of some specific phenomenon. However practical use of BI systems often is (1) descriptive and (2) diagnostic, but seldom (3) predictive or (4) prescriptive (Delen and Ram, 2018). The first two (1, 2) can be summarized as 'visualization', while the last two (3, 4) can be called 'fact-based decision making'.

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 analyzing information on for instance customer needs, organizational problems and improvement initiatives. The process from defining specific problems, choosing data to collect, analyzing 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).

At the heart of Quality Management lies Deming's System of Profound Knowledge (Deming, 1993):

- System view
- Understanding of variation
- Psychology
- Theory of knowledge

The Six Sigma methodology is built around these principles. Quality practitioners recognize that finding root causes to 'patterns in data' is a much more complicated and comprehensive task than just 'number crunching' of Big Data.

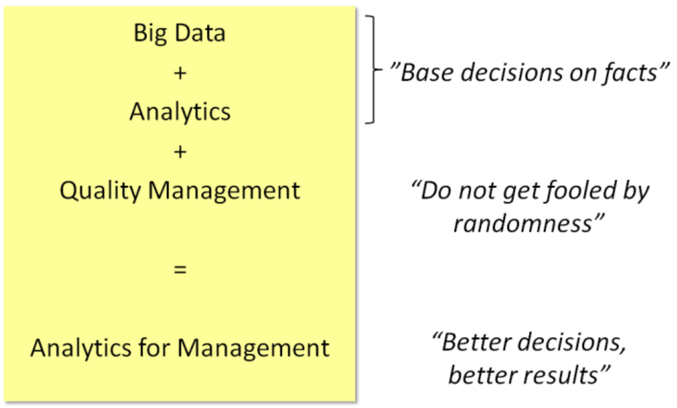
However, even with the success given by Six Sigma programs around the world, in the management rooms where the decisions are made, as Deming (1993) and Wheeler (2000) concluded, the principle of Fact-Based Decision Making has often been ignored, resting more on gut-feeling than facts. Even with *many data*, *i.e.* modern BDA, this is still the case (Cronemyr and Elg, 2014; Delen and Ram, 2018). As stated before, don't get fooled by randomness.

4.1 BDA – with and without feedback

The outcomes of BDA tools are often summaries and visualizations of many data, which is 'new' in the sense that you now have many data where you used to have few or none. However, there still is a tendency to view data in a rather traditional, *i.e. deterministic*, way. We see the potential of analyzing these data in a more *probabilistic* way, showing, and analyzing variation in data and its significant root causes; what we normally call QC tools. We also see a big potential in using the well-known QM tools for process-mining, data-mining and data-cleansing.

BDA on the one hand and QM on the other hand have until now mostly been two separate research streams. Cronemyr and Elg (2014) proposed to combine these into a new cross-fertilized research area investigating and developing new tools, procedures, and practices to utilize the full potential hidden in data, see Figure 1.

Figure 1. Analytics for Management



Source: Cronemyr and Elg (2014)

With ‘Analytics for Management’, *i.e.* BDA enhanced with QM tools and principles, Cronemyr and Elg (2014) suggested the potentials for many improvements. These improvements may be *e.g.*:

- Knowledge of variation,
- Customer focus,
- Process perspective,
- Root-cause analysis,
- Continuous improvement,
- Strategy deployment,
- Employee participation, and
- Feedback systems

The last item – Feedback systems – refers to systems for employees/process users to give feedback on the ways of working and the outcomes of the process, so process teams can continually update the process for increased customer satisfaction, employee satisfaction and business benefit. However, it comes with a risk, the risk of reacting too fast – tampering.

Another type of feedback, not addressed by Cronemyr and Elg (2014), is the ‘instant’ feedback of real data from the process. That could be called a ‘closed loop control system’. As known from basic control system theory, that may lead to out-of-control states and, in a worst case, divergence. That’s another type of tampering. This has not been addressed yet in research on BDA.

4.2 Fooled by randomness – Slow feedback leading to tampering

As previously mentioned, recent research indicates that although tampering is an old concept that is well known among quality management practitioners and researchers, it is 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. The term tampering is introduced in Deming's later writings (*e.g.* Deming, 1993) in conjunction with examples that less emphases on statistics. Following this broader view of tampering, Smeds (2021) suggests a contemporary view of tampering that highlights the usefulness of a more qualitative perspective on tampering as a compliment to the traditional quantitative perspective. The definition proposed by Smeds (2021) is *"[a] response to a perceived problem in the form of an action that is not directed at the fundamental cause of the problem, which leads to a deterioration of the process or the process output*" (p.47).

Two practices regarding BDA that risk leading to tampering is (1) over or under interpretation of the significance of certain data points and (2) incorrect identification of the causes of an event and consequently development of insufficient improvement efforts.

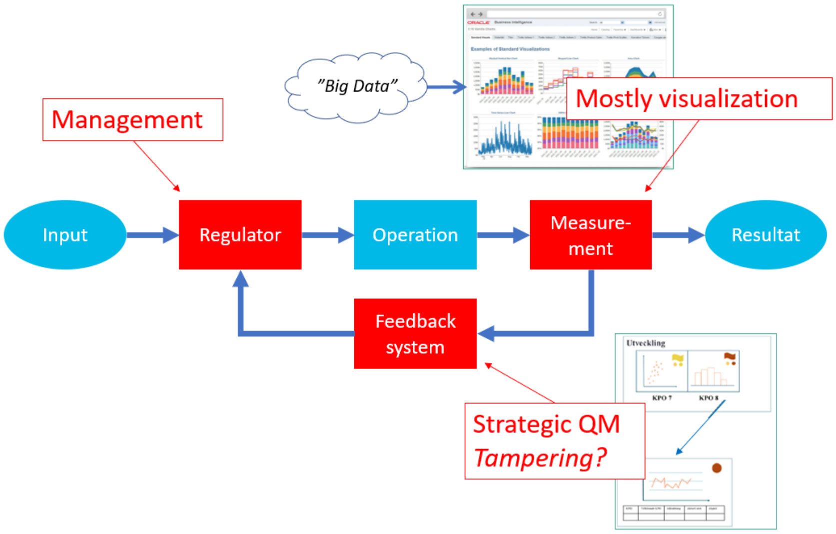
As previously pointed out, BDA provides new opportunities to make fact-based decisions. However, its effectiveness requires that data are analyzed properly. Understanding how data behave and the need for cause-and-effect analysis becomes especially important in the light of the research that indicates that humans do not have a natural understanding of statistics (Tversky and Kahneman, 1974) and variation (Coleman, 1999). Coleman (1999) even discovered that some students making an experiment on tampering found randomness to be unrealistic. Without this knowledge there is a risk that data that are considered significantly high or low without a reference to a process natural variation will be overestimated and lead to tampering (Deming, 1993).

A general challenge with data analysis is the time from data collection to analysis and consequent implementation of improvement efforts *i.e.* timeliness (Cai and Zhu, 2015). If data are analyzed *e.g.* 90 days after they were collected the data may be outdated and invalid (Cai and Zhu, 2015). Making decisions regarding an event that has arisen here and now based on old data will likely lead to uninformed decisions due to that data not reflecting the current situation. Timeliness of data analysis thus facilitates the ability to make timely responses and make reliable investigations of the causes of an event. Otherwise, the knowledge on possible causes or circumstances surrounding an event might be lost. Attempts to recall and explain past events may then be influenced by different types of biases such as availability bias or hindsight bias (Tversky and Kahneman, 1974). Such hindsight reflections may result in the wrong efforts being made to resolve that event. In addition, being preoccupied with what happened in the past may lead to events in need of attention in the present being neglected.

4.3 Process instability – Fast feedback leading to tampering

Another type of feedback, fast instead of slow, could be the 'instant' feedback of real data from the process, *i.e.* visualization of Big Data and some *following action*. Until now this has not been addressed in BDA research. As known from basic control system theory, a 'closed loop control system' may lead to out-of-control states and, in a worst case, divergence. That's another type of tampering. In Figure 2 below a *closed loop management control system* is presented conceptually. Without feedback it would be an 'open loop management control system', which is pretty much what all management systems look like today. However, a very slow feedback loop can be said to exist, *i.e.* 'learning by doing'.

Figure 2. A closed loop management control system

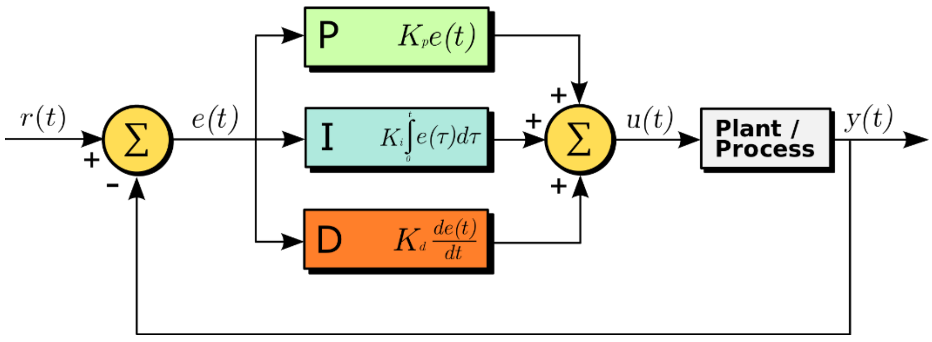


Source: Figure created by the authors

In Figure 2 comments about business management have been added to the description of the closed loop control system. 'Operation' is where the business processes are carried out. To control this, managers and operators are steering, controlling, and adjusting operations in the 'Regulator'. The results from the processes are measured in 'Measurement'. It could be an on-line measurements of Big Data, it could be off-line accumulated data with some frequency (day, week, month), or it could be the acquisition of personal knowledge and experiences (not really a measurement but can be treated as such, even though very slow). All of these measurements could be fed back to the 'Regulator' so inputs can be adjusted due to the difference of expected and measured results. It is often too slow as described above. Here Strategic QM/QC tools could be used (Cronemyr and Elg, 2014).

However, in the case of feeding back on-line measurements of Big Data, the 'Regulator' could go out-of-control (Capaci, 2019). It depends on the setting of the feedback system in the 'Regulator'. Basic control system theory describes three different types of feedbacks, see Figure 3.

Figure 3. A block diagram of a PID controller in a feedback loop



Source: https://en.wikipedia.org/wiki/PID_controller

The three different types of feedbacks in a PID regulator are:

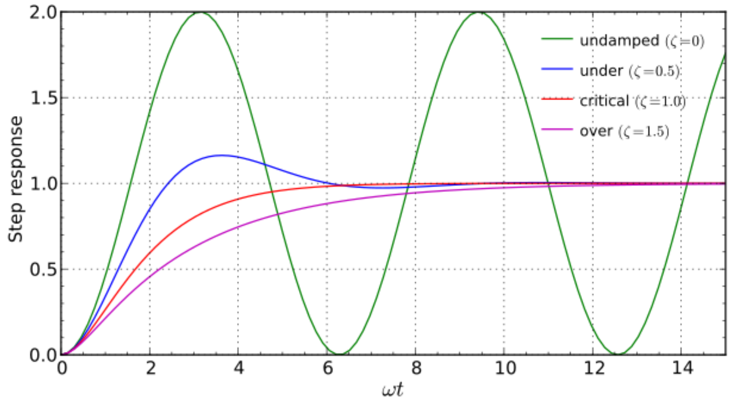
- P = A term Proportional to the difference of expected and measured results.
- I = A term that Integrates the differences over time.
- D = A term that Derivates the differences.

Depending on the settings of K_p, K_i and K_d, the system will have different characteristics. The full transfer function between input and response is often written (using Laplace transformation):

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad \text{eq. 1}$$

where ζ is the damping ratio and ω_n is the undamped natural frequency. Different response characteristics can be seen in Figure 4.

Figure 4. Step responses for a second order control system



Source: https://en.wikipedia.org/wiki/Control_system

We do not want the purple behaviour since it is too slow, and we clearly do not want the green behaviour since it leads to high fluctuations and possible out-of-control behaviour. The red and blue responses seem 'okay'; the blue response is a little faster but has some overshoot. The differences between the responses are given by the legend; the amount of damping.

Damping is the best solution to minimize tampering (however, not too little, and not too much), but what does it mean for a closed loop management control system? Clearly the damping is given by the settings of K_p, K_i and K_d, what do they mean for a closed loop management control system?

Maybe the following behaviors can help explain the responses:

- P = Managers compare ‘where we are’ to ‘where we want to be’.
- I = Managers consider the level of differences ‘we have had’ recently.
- D = Managers look at the ‘slope of the curve’ showing the differences and let that influence the level of control.

We do not know much about any of these three types of tampering. However, we suppose classical type I and II errors are typical ‘*P behaviors*. We have seen ‘learning from previous tampering’ behaviors that look like ‘*I behaviors*’ (Smeds, 2022). Finally, the increasing clock-speed of processes in combination of BDA without root cause analysis may lead to ‘*D behaviors’*. That is what we would like to research more about, hence the following research idea is suggested.

4. A research idea

There are no real results in this paper. However, we have given the background and rationale of conducting more research in this field. We propose to investigate:

- How can different types of feedback be classified as P, I, and D behaviors?
- Which of these behaviors could be classified as tampering, and should be avoided?
- How could P, I and D tampering be avoided, *i.e.* how could K_p, K_i and K_d in the closed loop management control system be set to minimize tampering?
- 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 (1) interpreting data correctly and, (2) making the correct actions?
- When is tampering (1) a cause and (2) when is it an effect?
- To what extent and under which circumstances could traditional QM/QC tools be used in strategic (i.e. planning), tactical (i.e. preparing and improving), and operational (i.e. real-time response) management system, thereby reducing tampering?

More questions can of course be added. It is our intention to start up research initiatives in cooperation with our contacts in industrial businesses as well as in public organisations. We also need more academic cooperation in this area. Do you want to participate?

References

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

Bergmark, W. and Linderstam, J. (2021), *Bättre beslutsunderlag med återkoppling från processer – Digitalt återkopplat styrsystem för ledning och styrning*. Linköping University, Management and Engineering, Master thesis, LIU-IEI-TEK-A--21/04106--SE.

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.

Cai, L. and Zhu, Y. (2015), The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002.

Capaci, F. (2019), *Adapting Experimental and Monitoring Methods for Continuous Processes under Feedback Control: Challenges, Examples, and Tools.* Doctoral dissertation, Luleå University of Technology.

Chowdhury, T. and Sandén, L. (2015), *Analytics for Management – En modell som beskriver framtagandet av ett beslutsunderlag där rätt mätetal visualiseras på rätt sätt utifrån en formulerad strategi.* Linköping University, Management and Engineering, Master thesis, LIU-IEI-TEK-A--15/02144--SE .

Coleman, D.W. (1999), Adapting Deming's funnel experiment to a content-specific area. *Simulation and Gaming*, 30(1), pp. 8-19.

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., Fundin, A. and Wemme, L. (2022), Change Driven Process Management. *In the proceedings of EurOMA Conference,* Berlin, Germany.

Davenport, T., Barth, P., and Bean, R. (2012), How 'Big Data' is Different. *MIT Sloan Management Review*. Fall 2012, Vol. 54, No. 1., pp. 22-24.

Delen, D. and Ram, S. (2018), Research challenges and opportunities in business analytics, *Journal of Business Analytics*, 1:1, 2-12, DOI: 10.1080/2573234X.2018.1507324.

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

Jordan, J. and Lin, D. (2014), Statistics for Big Data: Are Statisticians Ready for Big Data?, *ICSA Bulletin,* January 2014, Vol. 26/1. Controversial issues, pp. 58-65.

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

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. and Kruschwitz, N. (2011), Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*. Winter 2011, Vol. 52, No. 2., pp. 21-31.

Liu, S. (2014), Breaking Down Barriers. What quality professionals need to know to survive and thrive in the age of big data. *Quality Progress*. Jan. 2014, 47, 1, pp. 18-22.

MacInnis, D.J. (2011), A framework for conceptual contributions in marketing. *Journal of Marketing*, 75(4), pp. 136-154.

Magnusson, C. and Nordlund, E. (2020), *Digital Tools For Analysis And Improvement – Incorporation Of Digital Tools In Improvement Processes*. Linköping University, Management and Engineering, Master thesis, LIU-IEI-TEK-A 20/03736--SE.

Mayer-Schönberger, V., and Cukier, K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Houghton Mifflin Harcourt.

McAfee, A. and Brynolfsson, E. (2012), Big Data: The Management Revolution. *Harvard Business Review*, October 2012, pp. 2-9.

Sehlin, D., Truedsson, T. and Cronemyr, P. (2019), A conceptual cooperative model designed for processes, digitalisation and innovation. *International Journal of Quality and Service Sciences* Vol. 11 No. 4, 2019 pp. 504-522.

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

Smeds, M. (2021), Deming's tampering revisited: definition and future research agenda. *International Journal of Quality and Service Sciences*, Vol. 14 No. 5, pp. 37-51.

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.

Tversky, A. and Kahneman, D. (1974), Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), pp. 1124-113.

Wheeler, D. (2000), *Understanding Variation – the Key to Managing Chaos*. (Second edition). SPC Press, Knoxville.